TONY TSAI

GEOMETRY-AWARE AUGMENTED REALITY FOR REMOTE COLLABORATION

GEOMETRY-AWARE AUGMENTED REALITY FOR REMOTE COLLABORATION

TONY TSAI

A thesis submitted in partial fulfilment
of the requirements for the degree of
Master of Engineering
College of Engineering
University of Canterbury

December 2015 – version 4.1

Tony Tsai: *Geometry-Aware Augmented Reality for Remote Collaboration,*
Master of Engineering, © December 2015

SUPERVISORS:
Dr. Steve Weddell
Prof. Mark Billinghurst

Dedicated to my parents, for their unconditional love and support
and to my supervisors, whose indispensable guidance and support
helped me immeasurably during crucial moments in this project.

ABSTRACT

The task of repairing or maintaining complex machinery in a remote collaborative environment can be very challenging for the involved parties, both on- and off-site. Augmented Reality can offer a more immersive experience in this situation for both users. Using both sparse and dense Simultaneous Localization and Mapping (SLAM) techniques, two different systems for remote collaboration were developed. Annotations are provided for the on-site user, and the task environment is re-created for the off-site user. These system allows complex remote collaborative tasks to be conducted more effectively. The main contribution of this research was the investigation and application of various technologies and algorithms to build working prototypes to satisfy a set of requirements.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Matthew Tait, Tony Tsai, Nobuchika Sakata, Mark Billinghurst, and Elina Vartiainen. "A Projected Augmented Reality System for Remote Collaboration." In: *ISMAR2013 - International Symposium on Mixed and Augmented Reality*. 2013.

Tony Tsai, Mark Billinghurst, Elina Vartiainen, Fredrik Alfredsson, Jonas Bronmark, and Mathew Tait. "Method and Data Presenting Device For Assisting A Remote User To Provide Instructions." Patent PCT/EP2014/059558 (Europe). 2014

ACKNOWLEDGEMENTS

This thesis was funded by scholarships from Asea Brown Boveri Group (ABB) and The University of Canterbury.

I would like to thank both of my supervisors, Dr Steve Weddell and Prof. Mark Billinghurst, for their help and guidance through this process.

I would like to also acknowledge the work of Seungwon Kim, whose work was used as a starting point for my own, and to Mathew Tait, who contributed to the work for the projection system.

I would also like to thank the people at ABB Corporate Research, Västerås Sweden, and the SECR group for their support which made this project possible.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

ABB	Asea Brown Boveri Group
API	Application Programming Interface
AR	Augmented Reality
DOF	Degree of Freedom
DPR	Digital Pixel Reflection
FOV	Field of View
GPU	Graphics Processing Unit
GUI	Graphical User Interface
KinFU	KinectFusion [25]
HITLab NZ	Human Interface Technology Laboritory New Zealand
HMD	Head-Mounted Display
HUD	Heads-Up Display
IMU	Inertial Measurement Unit
ICP	Iterative Closest Point
MUX	Multiplexor
OS	Operating System
POV	Point of View
PTAM	Parallel Tracking and Mapping [18]
SLAM	Simultaneous Localization and Mapping
UI	User Interface
VoIP	Voice-over-Internet Protocol
UML	Unified Modeling Language

TERMS AND SYMBOLS

Pose	Position in space of the SLAM agent
On-site	Remote location where the work is being performed
Off-site	Any location remote from the work environment
Client	Each on-site unit
Server	The controlling off-site system
Technician	On-site personnel who directly carrys out the work
Specialist	An expert with specific skills or knowledge to effect repairs on a specific piece of machinery.

Remote Expert An off-site specialist who is using a remote collboration system to effect repairs without being there.

$R^3_{(x,y,z)}$	Coordinate axis in the 3D World frame
$R^3(u,v)$	Projection point axis for camera image plane.
W	World co-ordinate frame
C	Camera co-ordinate frame
E_{AB}	Transformation Matrix from co-ordinate A to B

1

INTRODUCTION

As our world becomes more and more complex, so do the machinery and systems we use in our day-to-day lives and various industries. This increase in complexity means that the regular maintenance and repair of these machines require personnel with more training and specialization. The sheer diversity of machinery and specialization required has meant that there may often only be a handful of engineers experienced enough to perform the required tasks.

With machines operating in remote locations and the increase in globalization of machine manufacturers, the specialists required to perform a task may not be in the same site or even in the same country. These specialists will have to travel to the site to effect repairs which can involve being flown half way around the world. The extended time taken before repairs are completed can cost a company dearly in terms of down-time and travel costs.

The locations of personnel and work sites in a typical plant is shown in Figure 1. To outline the problems and complexity that may arise, a typical example of a mock scenario is given as follows:

> An important piece of control equipment for an oil drilling platform breaks down. The platform is located 40 Km from shore and can only be accessed via a boat or helicopter. The technicians on the platform are unable to repair the equipment. A call is made to the manufacturer of the equipment, who puts one of their specialist engineers on to help fix the problem. Numerous calls and e-mails with pictures are exchanged to no avail. Finally, it

is decided to send the specialist out to the rig. Boarding a rig is a very complicated procedure which requires checks on personnel and only certified equipment is allowed on board. After lengthy delays, the specialist finally arrives on the rig and is able to inspect the machine. They find that repairs could have been carried out by the original technician if the true condition was able to have been conveyed remotely and the repair sequence could have been actuated effectively.

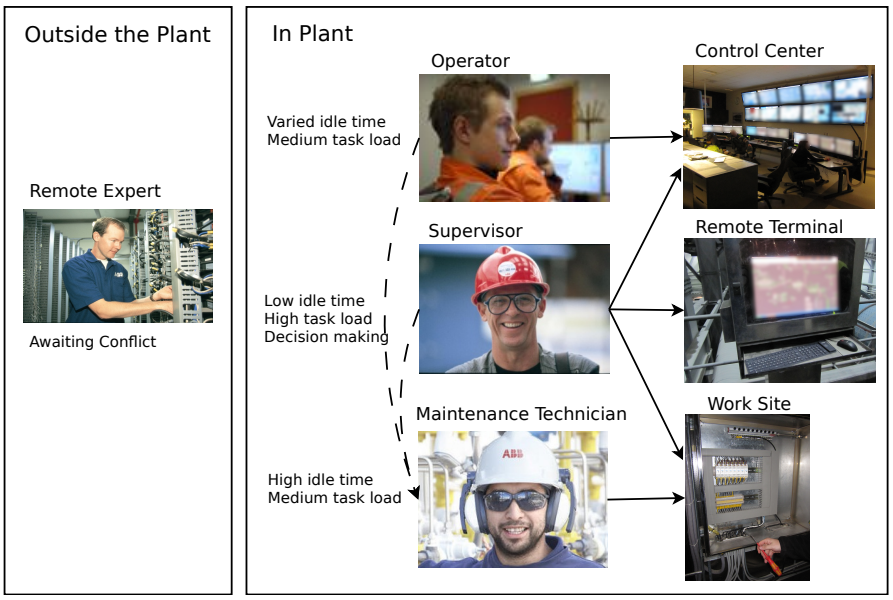


Figure 1: Mock remote collaboration scenario with players and locations. The current flow of instructional information is shown with a dotted line. The solid arrows show the locations in which each player works.

1.1 OBJECTIVES

To avoid the problems described in the previous section, effective remote collaboration techniques are investigated in this work between the Human Interface Technology Laboritory New Zealand (HITLab NZ) and ABB [21]. Remote collaboration allows two parties, who are separate from each other, to work on a common task using a variety of communication tools. This allows an inexperienced technician, who is

on-site, to collaborate remotely with the specialist to fix the problem, removing the need for the specialist to be sent to the on-site location.

Current methods for this type of remote collaboration typically involve emails, phone calls, or real-time video communications such as Skype[1]. There are, however, many limitations with these current protocols for remote collaboration, which can be exacerbated by a complex repair situation. These limitations exist primarily due to two types of disconnects. The first disconnect is the misunderstanding that the specialist may have of the on-site situation. The second disconnect occurs from an error in the on-site technician's understanding of the specialist's repair solution. These are at their worst when only voice communications are used. The use of e-mail allows different media types to be transmitted but can cause other problems due to its long round-trip response time. The use of real time video communications gives a better sense of the situation for the specialist, but problems still arise with the conveyance of solutions from the specialist due to the inability to point and gesture.

To facilitate collaborative repairs in a fast and efficient manner, a set of guiding features or targets for our system were established. These recommendations include:

INCREASED SITUATIONAL AWARENESS: The ability for the specialist to fully understand the situation on-site is vital. As much real-time information about the site should be given as possible. Extra peripheral and auxiliary information, beyond visual and auditory information, should also be included.

EFFECTIVE ACTUATION: Once the situation is fully understood by the specialist and appropriate repair actions are decided upon, those actions must be effectively performed on-site. The specialist must be able to direct the technician and convey his/her message.

1 Skype is a face-to-face direct Voice-over-Internet Protocol (VoIP) communication systems which allows the real-time transmission of voice and video between multiple parties. http://www.skype.com/

EASE OF SETUP: Many current technologies, which offer the auxiliary information sought by a technician, have demanding requirements in terms of equipment, bandwidth or computational power. The provided system should be easy to set up and require minimal equipment.

INTUITIVE INTERFACE: The feeling of being on-site must be as strong as possible for the specialist. Conversely for the technician, the feeling of having a support person present must be as strong as possible while the work is being carried out. By maximizing these feelings, the constructed system would be more intuitive to operate.

Two separate systems were developed during the course of this project to investigate the effectiveness of the different options. To decide on and deliver the functioning prototypes, background research was conducted into the current state of technology and systems that were available at the time. This preliminary research is detailed in Chapter 2. It discusses why these current tools, used in both industrial and research laboratories, are less than optimally effective and it also provides a road-map for future development. Chapter 3 then first discusses the framework on which both systems are build and covers the work that is common for both systems. Chapter 4 discusses the first system which was developed to overcome the shortcomings of the prototype system discussed in Section 2.1.1. This head-mounted system implements a robust Simultaneous Localization and Mapping (SLAM) algorithm to increase the ease of understanding annotations while simultaneously building an interpolated 3D map of the world. Chapter 5 then discusses a projection based system which projects of annotations directly onto the real world using a micro-projector. This system was built to compensate for the shortcomings, regarding the technician's perception and freedom of movement, of the first system. The motivation, implementation, novel work, and results for each system are discussed in more detail in their respective chapters.

2

BACKGROUND

Given the current problem of effective remote collaboration detailed in the previous chapter, a background investigation was conducted into the current state of technology and systems currently being used for remote collaboration.

2.1 CURRENT SYSTEMS

Most current forms of remote communications are audio and video conferencing tools which are typically designed to support face-to-face communication. The ABB proposal discussed in Chapter 1, however, requires more effective task-space collaboration. When face-to-face systems are used in task-space collaborations, the results are far from optimal. The time taken to performs these task can be much longer with more errors being committed [7, 19, 4, 1].

The focus of this research will be on communication systems that are task-space oriented. Many current task-space communication systems operate in a common manner. The task-space can be shared between the two users typically using video or still images. An example of a task-space collaboration system is shown in Figure 2. Annotations are then overlaid for each user to view. Most of the systems differ in three aspects: 1. how they implement the SLAM algorithm; 2. how augmented information is displayed to each user; and 3. how the users interact with the system and virtual information. The systems listed below cover a range of different implementations for these three aspects of a remote collaboration system and given as examples:

POELMAN ET AL. 2012[31] This system uses the PTAM [18] algorithm along with a dual camera differential imaging system to perform SLAM tracking. This allows an on-site technician the ability to roam free while still viewing annotations made by either the local or remote party. Gesture recognition is implemented to allow the local user to input information into the system.

DRAW OVER VIDEO ENVIRONMENT (DOVE) [28] This system supports the drawing of annotations on a live video feed of a task environment. Permanent annotations, temporary gestures, and live cursors can be shared between the two users. These separate modes of drawing control the persistence of information and helps to avoid the saturation of the work environment with augmented information.

ANNOTATING WITH LIGHT [29] A fixed camera and laser projector are used to display annotations drawn by a remote expert. However, due to the lack of depth data, projections on non-planar surfaces are subjected to an offset error proportional to the distance from the target plane.

TELEADVISOR [9] TeleAdvisor is a versatile augmented reality tool for remote assistance. A sample illustration of the operation of TeleAdvisor is shown in Figure 3. Annotations are projected onto the world using a laser galvometer that is fixed, along with a camera, to a remotely controlled arm.

PLATONOV ET AL. [13] Anchored annotations are shown on a handheld tablet with a fixed camera. Pre-loaded information, in the form of a CAD model, along with edge detection, is used to aid in the tracking of pose. Predefined complex annotations can then be shown in the environment on the tablet.

Besides these systems discussed above, a separate system was also being developed at the HITLab NZ to facilitate the task of remote

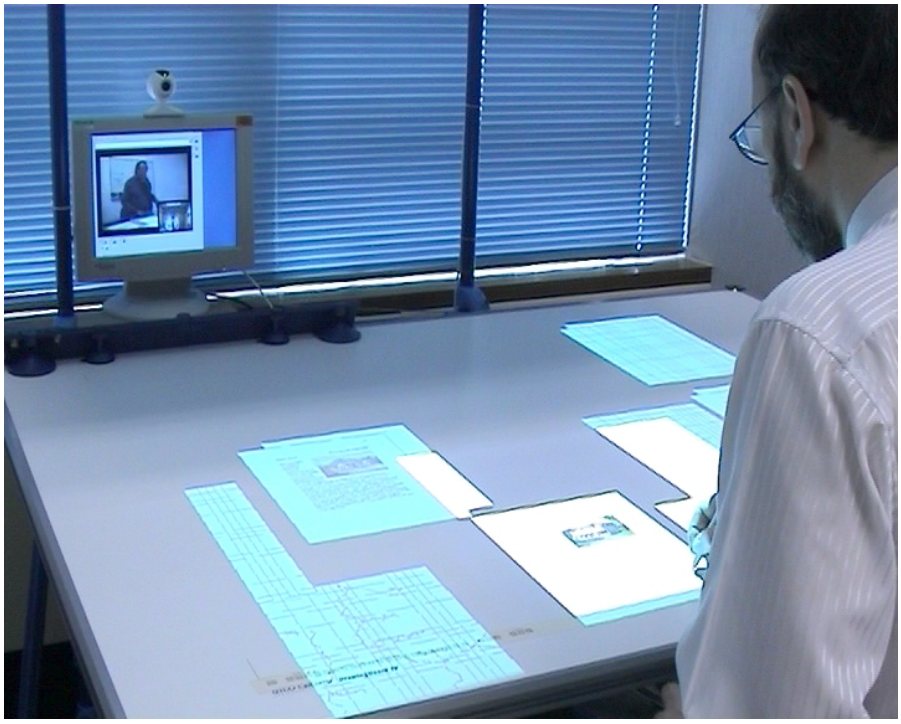


Figure 2: An example of a task-space collaboration system where the focus for both users is the desk in front of one of them. The system shown is called the Escritoire by Ashdown and Robinson [2].

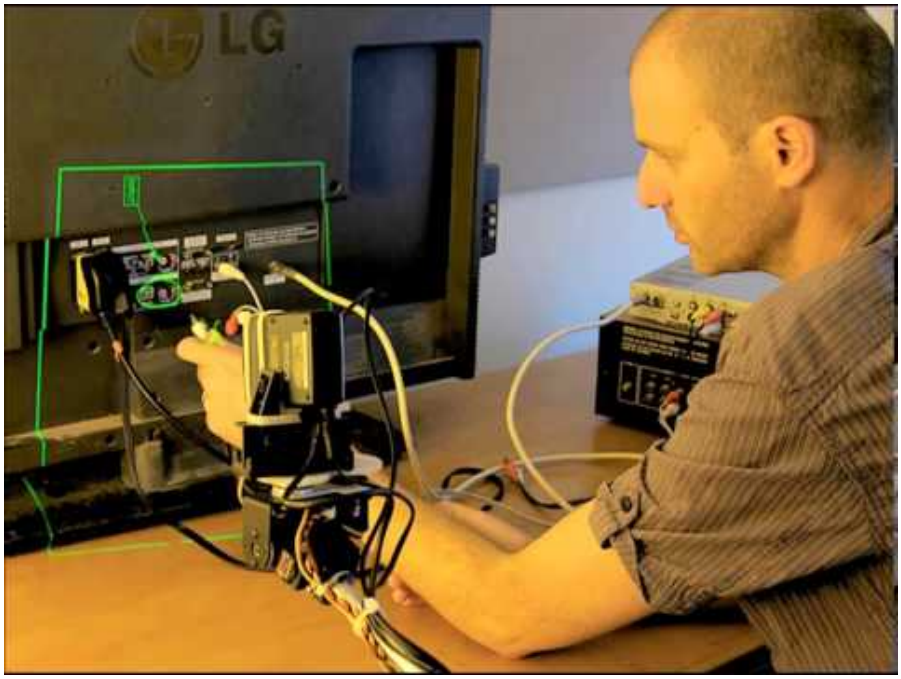


Figure 3: Laser projection of annotation data using the TeleAdvisor system developed by Gurevich et al. [9].

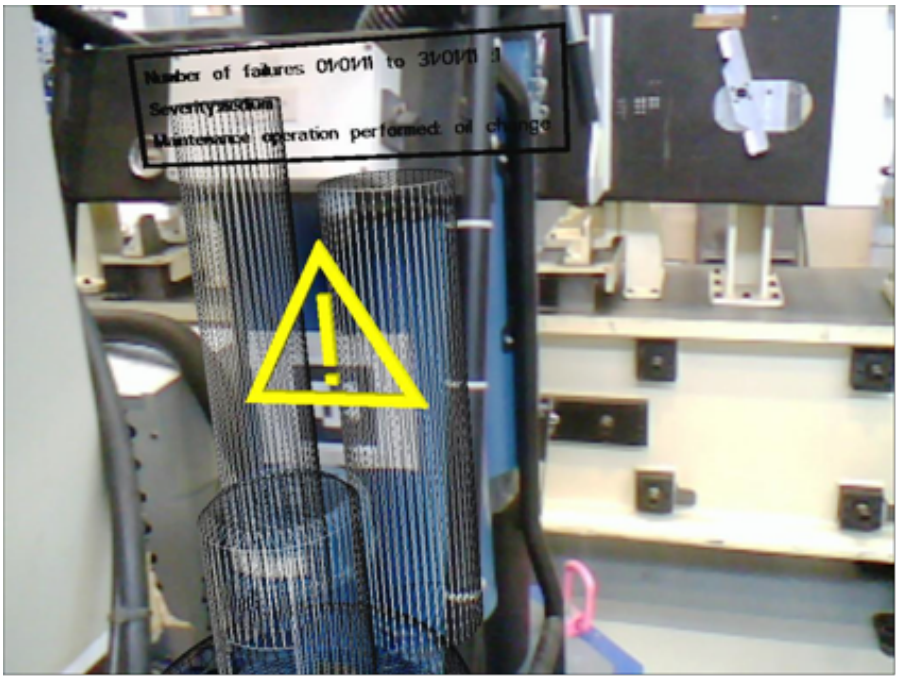


Figure 4: A sample of augmented information shown on real world objects. Cylinders and exclamation sign digitally added to the scene using ManuVAR by Toslin et al. [37].

collaboration and is discussed in more detail in Section 2.1.1. New technologies which will greatly influence how a remote collaboration system is used and operated, and these are discussed in Sections 2.2 and2.3.

2.1.1 *Remote Collaboration Group at HITLab NZ*

Task-space orientated systems are much less common, have limited industry adoption, and tend to be expensive and overly complex to operate, so research into alternative systems is being conducted at the HITLab NZ. When the research for this thesis started, work on a different novel system had already been started by another member of the HITLab NZ, Seungwan Kim.

The developed system consisted of an Android device[1] that is connected to a PC via a network connection as shown in Figure 5. The on-site technician uses the Android device to stream video from the task space to the off-site specialist. This video link is the main mode of communication with annotations in the form of simple drawings and text, being shared between the users.

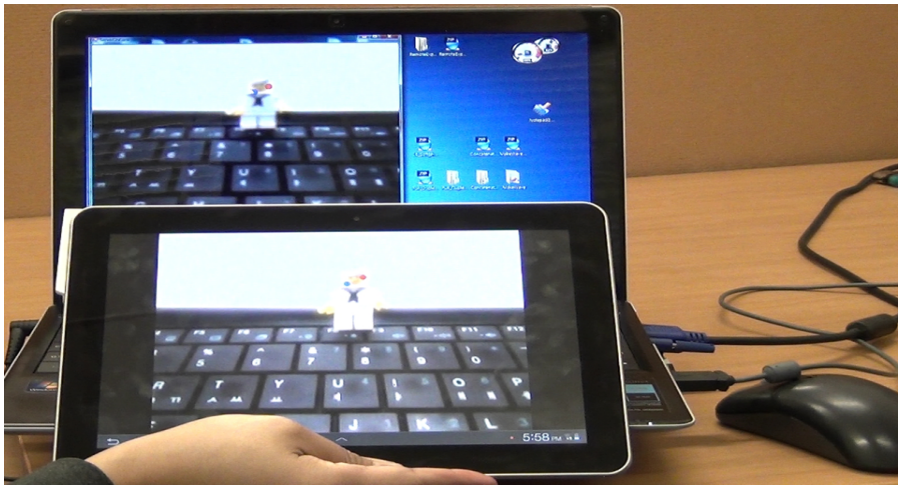


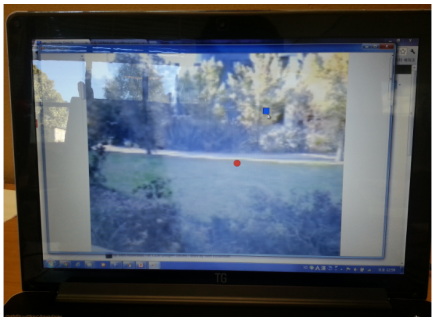
Figure 5: Hand-held Android based remote collaboration unit. The camera from the table is streamed to the Laptop. On both screens you can see both a blue and red cursor. Each cursor is controlled by their respective device.

Operation of the system

Two modes of operation are available to users of the system including a live stream mode and gallery mode. In the default live streaming mode, video from the on-site user's Android Tablet is sent to the off-site user. Two cursors are overlaid on the video stream for both users. One cursor is controlled by the on-site user and the other by off-site user. To transition into the gallery mode of operation, shown in Figure 6d, each user can take a screen shot of the live video and draw annotations and text on it before sending it to the other user. Previously sent pictures can also be retrieved from a gallery.

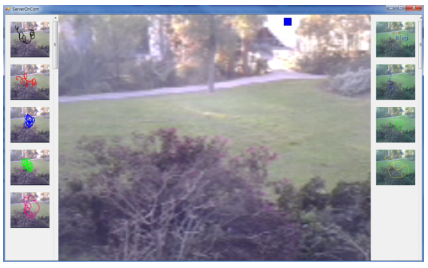
1 The software is compatible with any mobile device running the Android Operating System (OS) version 2.3+. A video-capable camera and network connection are required to stream the live video to an off-site PC.

(a) The Android tablet showing the live view of the scene.



(b) The view from the off-site user of the stream with respective cursors.

(c) Annotation made from a screen capture of the streaming video.



(d) The saved gallery of sent images. Images on the right from the on-site user and left from the off-site user.

Figure 6: Operation of the Android streaming system showing both modes of operation, cursors, and annotations.

Review

After utilizing the system developed at the HITLab NZ for a period of time, some distinct advantages and disadvantages of the system became apparent.

ADVANTAGES

- The system is very robust and reliable.
- It is readily deployable with minimal extra cost.
- The development time for the project is very modest.

DISADVANTAGES

- The technician's hands are occupied holding the device.
- It requires the device to be held steady for the cursors to work well.
- Lag in the streaming of the video greatly reduces the usability of the cursors.

Along with a review of current systems developed by Seungwan Kim, other emerging technologies were also investigated to see how they may increase the effectiveness of developed systems.

2.2 NEW TECHNOLOGIES

The introduction of new technologies has also provided a unique opportunity to develop novel collaborative systems. These technologies affect how information is captured, presented and interacted with. Selected technologies are discussed here based on their viability to facilitate remote collaboration.

2.2.1 *Head Mounted Displays*

Since the mode in which information is displayed to the user greatly affects many factors, including usability and experience, different types of displays are investigated. For our application, HMDs are a good alternative to typical desktop monitors and hand-held displays. HMDs allows users to view information while minimizing the restriction to their movement and freeing their hands for other tasks.

There are many aspects of a HMD that affect the user's experience. These include Field of View (FOV), resolution, screen opacity, and focus distance. The FOV of a HMD determines what percentage of the user's vision is covered by the display. Typically higher FOVs allow for a much more immersive experience. Along with the FOV, the resolution determines the density of pixels that is displayed. Again as with the FOV, higher resolutions contribute to a more immersive experience. Screen opacity determines how much natural light comes through the display. Displays with 100% opacity rely on a camera to record and display the real world. Displays with partial opacity allow the user to view the world directly with the augmented information superimposed on top. Partially opaque displays can remove feelings of disconnection to the real world and in some cases nausea. However, they do not display the virtual annotations as vividly as opaque displays. The focal distance of a display is relevant only for partially opaque displays. It determines at what distance the user must focus their eyes to view the virtual information clearly. Some displays can have their focal distance adjusted. However, in an environment where the distance to the workpiece can vary, adjusting this can be troublesome. Newer displays can also feature infinite focus distances that stay sharp no matter what distance the user is focused on by using laser projection.

All HMDs can also be separated into one of two categories listed below:

MO - MONOCULAR: These displays cover only a small portion of users natural FOV and can be either an opaque or semi-translucent display. An example of this type of display is shown in Figure 7.

ST - STEREOSCOPIC: These displays cover both eye's of the user and occludes the entire FOV. Because of their large coverage, they tend not to be translucent. An example of this type of display is shown in Figure 8.

Each type of display can also be one of two optical transparencies:

O - OCCLUDED: These displays completely obscure all natural light that would normally pass through to the user. Information about the world is generally captured with a RGB camera and shown on the screen with augmented information overlaid on top.

S - SEE-THROUGH: These displays allow natural light to pass through to the user. Any changes of the display relative to the user's eye will cause a shift of alignment between the displayed annotations and the view of the real world. Causing annotations to become misaligned.

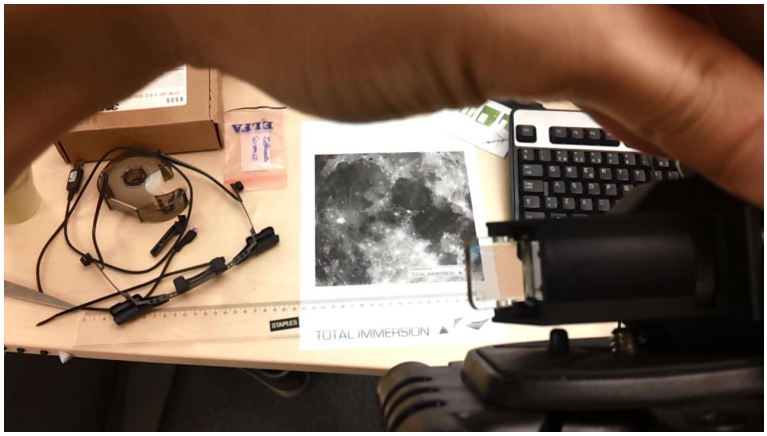
Table 1 lists a selection of HMDs that are currently on the market that were investigated for this work, including the category to which they belong.

Emerging technologies with ultra-wide FOV displays are being developed at the time the research for this thesis was being conducted. The Occulus Rift[2] is the earliest model of these new ultra-wide FOV displays. This technology is significant because the ultra-wide viewing angle gives an unprecedented level of situational awareness while presenting augmented information in the entire FOV. This can be compared with other non-see-through displays whose narrow FOV reduces the user's situational awareness, critical in hazardous work-

2 https://www.oculus.com/

Table 1: Comparison of current HMDs

HMD	CATEGORY	RESOLUTION	FOV
Motorola Golden-I	MO, S	800x600	–
Vuzix Tac-Eye	MO, O	852x600	30°
Smart Vision Laster	ST, S	800x600	40° – 30°
Laster G1	ST, S	800x600	40° – 30°
Rockwell Collins MicroView	MO, S	800x600	29° – 21°
Laster Pro Mobile Display	MO, S	800x600	40° – 30°
Lumus' OE-32	ST, O	1280x720	40°
Vuzix Smart Glasses	ST, O	1280x720	30° – 50°
Brother Airscouter	MO, S	800x600	22°



(a) Live view of the world with the Brother Airscouter in the bottom right.

(b) View of the augmented information through the glass of the Airscouter.

Figure 7: Brother Airscouter (MO,S) displaying information through the 50% transmissive glass; (a) live view; (b) augmented view.

Figure 8: The Vuzix Smart Glasses (ST,O). The view of the world is suppled by a head-mounted RGB camera and viewed on the screens in front of each eye.

ing environments. This type of ultra-wide viewing angle displays was released after the work for this thesis had been done so their use was not considered for implementation in the project. They are mentioned here for completeness and future reference.

2.2.2 *Projectors*

As discussed previously with the HMDs, the mode for displaying information greatly affects the experience and efficiency of the system. Because of this, other means of displaying information are investigated. Projectors provide an interesting alternative to view information. Rather than having the information on a screen in front of your eyes, the information is projected onto the real world. Several projector technologies are pivotal in allowing projectors to be used in this type of dynamic application:

GALVANOMETER LASER PROJECTION: A single laser beam is deflected on a computer-controlled reflective surface to produce graphics. This allows for an image with infinite focus, removing the need for the user to adjust focus for a sharp image. This also increases

the brightness and contrast of the display. An example is shown in Figure 9.

DIGITAL PIXEL REFLECTION: Digital Pixel Reflection (DPR) operates in a similar manner to Galvanometer Projection. However, instead of a single reflective surface, an image is formed by individual pixels, each controlled by an individual mirror which can be in either an 'on' or 'off' state. This allows for higher contrast to be projected compared with traditional LCD displays. This comparison is shown in Figure 10.

SOLID STATE BULBS: This type of bulb include LED and laser light. The development of this type of bulb allows projectors to fit in a smaller package while also consuming less energy with a relative increase in luminance compared with current filament-based bulbs. Laser LEDs also provide the benefit of infinite focus, which was also discussed earlier.

SELECTIVE LIGHT POLARIZATION: By polarizing the light of two projectors orthogonally, unique information can be presented to each eye. This can give the user the impression of 3D information with the use of polarizing glasses.

These newly released projectors continue to come in smaller packages with reduced weight while still offering increases in brightness as shown in Table 2. Projectors are no longer relegated to use in darkened rooms and in a fixed position. This allows them to be used in mobile situations and in more varied lighting conditions, and they can now be considered for use in our application as an alternative to HMDs.

2.2.3 *Depth Cameras*

While there have always been a variety of depth cameras available, the introduction of the Kinect camera[23] has lowered the bar for

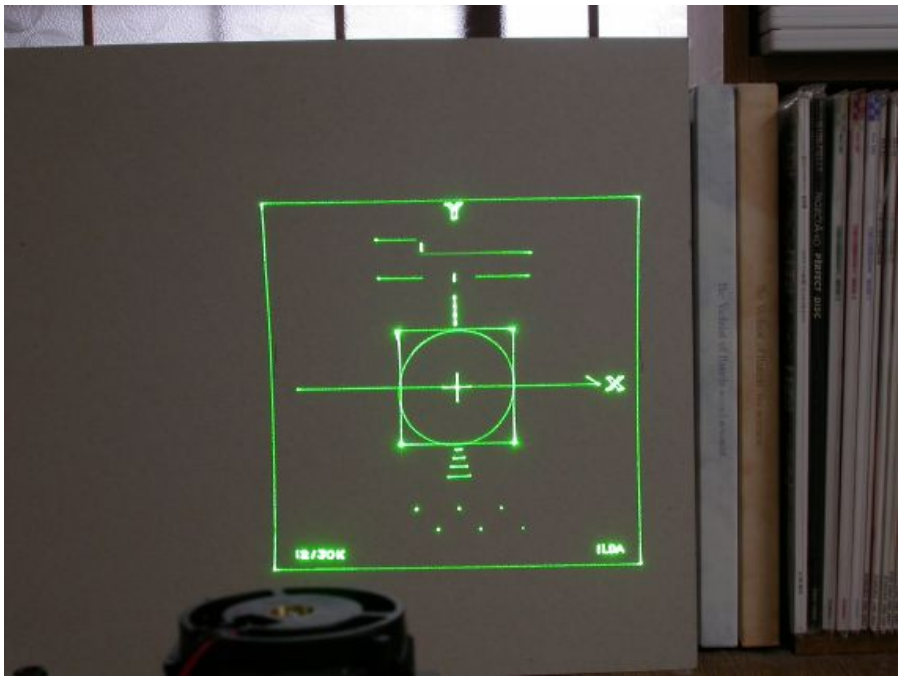


Figure 9: A projected image using a galvanometer-controlled laser shown on a flat surface. The projected information can be viewed comfortably in a well-lit room

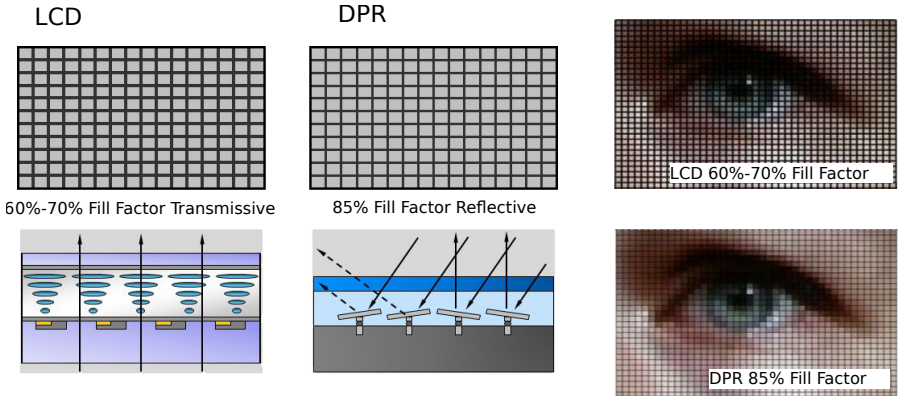


Figure 10: A comparison of LCD vs DPR projector technology. DPRs allows for higher contrast and reduced "screen door effect".

Table 2: Brightness of projectors in comparison with their weight for solid state and laser bulb projectors.

	BRIGHTNESS (LUMENS)	WEIGHT (KGS)
Pico Projectos	200	0.3
Pocket Projectors	500	1.1
Portable Projectors	2000	2

implementation, offering low hardware costs and readily available user libraries. The use of depth cameras allows the surroundings to be mapped without the use of computationally expensive algorithms to derive three dimensional data. Kinect uses a spatially encoded IR pattern to determine depth information. An example of this light pattern is shown in Figure 11. Advantages and disadvantages of this technique are discussed below:

Figure 11: The infra-red pattern projected by Kinect which is used to determine depth data from the scene. [35]

Spatial encoding sensor

Advantages:

- It has a lower reliance on the capabilities of the feature extraction algorithm.
- Varied lighting conditions do not have a marked effect on it's ability to function.
- Ability to capture information in a single frame compared to other sensors such as structured light sensors.

Disadvantages:

- Compared with other technologies such as time-of-flight cameras, is is more susceptible to noise.
- Infrared implementations will only work indoors out of direct sunlight.
- Large amount of data can become strenuous on communications protocols.
- Low resolution depth data 640×480 pixels is generated from the 1600×1200 infra-red sensor. [12]

Further research and analysis of the Kinect camera, done by Khoshelham shows that the effective range for accurate distance mapping is between 0.8 m to 3.8 m [16]. Khoshelham's paper concludes that the Kinect sensor does not contain large systematic errors compared with laser-scanned data. The error increases exponentially with distance, with a maximum error of 4 cm at the furthest distance. More detailed information about the implementation of the Kinect camera is shown in Section A.2.1 of the Appendix.

2.3 SIMULTANEOUS LOCALIZATION AND MAPPING

A shared 3D view and Augmented Reality (AR) elements overlaid on real objects have been shown to aid remote collaboration [10]. To achieve this, an effective SLAM algorithm is needed to simultaneously map the surroundings of the on-site user and track the camera position relative to those surroundings. The SLAM algorithms that are currently being considered can be divided into two categories: sparse mapping and dense mapping. Sparse mapping algorithms such as those proposed by many different researchers [18, 20, 17] perform localization based on a limited number of extracted features from the scene. Due to this, they can be implemented in real time

on relatively modest consumer-grade hardware. New dense mapping algorithms such as Parallel Tracking and Mapping [18] (PTAM) developed by NewCombe et al. [24, 25] perform point tracking on a per-pixel basis and produce much more detailed models. These algorithms can also be run in real-time but require the use of massive-parallel-processing such as Graphics Processing Unit (GPU)s. Using the aforementioned methods to generate a map, users can share spatial details with remote parties. This has been shown to improve remote collaboration [5].

2.4 IMU FOR TRACKING

An Inertial Measurement Unit (IMU) can provide a high rate of accurate acceleration data, both linear and rotational. In addition to this, integrated magnetometers can provide relatively accurate ground-based truth data. Their low hardware costs and almost ubiquitous inclusion in devices such as smart phones and HMDs make them a prime candidate for data fusion when estimating poses. The list bellow covers most of their advantages as well as disadvantages compared to vision based pose estimation.

Inertial sensors

Advantages:

- The self-contained nature of such devices removes the reliance on other hardware.
- Their ability to supply samples at a high rate (in the KHz range) ensures a steady stream of data.

Disadvantages:

- The need to use a double integral to compute displacement can introduce significant error.[3]

3 Initial values can be supplied from vision system guesses.

- The self-contained nature of these devices does not provide external "ground truth"

Vision Sensors

Advantages:

- Measurements from the viewing position can minimize the visual alignment error.

Disadvantages:

- Repeated patterns in the environment can cause a loss of tracking.
- They can have high computational cost
- They are susceptible to varying light conditions
- They are highly dependent on the ability of the feature extraction algorithm.

Many papers have shown that IMUs can improve the accuracy, robustness, or computation requirements of pose estimation [11, 39, 40].

2.5 REVIEW

After evaluating the current system at the HITLab NZ and reviewing other systems listed in Section 2.1, certain areas for improvement had become apparent. The lack of a robust SLAM algorithm for the HITLab NZ system made annotations hard to understand with any movement of the hand-held camera. The system by Poelman et al. [31], while effective as a remote collaboration system, could be ungainly to use with the amount of hardware attached to the user's head. The DOVE system [28], removes the hardware attached to the user's body, but lacks the ability to capture depth information from the scene, which caused inaccuracies in the projected annotations.

Based on this, a few principles for improving on remote collaboration system are established.

The system should have minimal impact on the user's mobility, dexterity, and perception of the environment. As much as possible, hardware attached to the user must be minimized without changing the system's ability to function. A stable localization algorithm will greatly increase the understanding of annotations. For the off-site user, the concept of situational awareness of the task space is most important. The inclusion of auxiliary information about the site, including 3D data of the environment and location of the on-site technician, will give them a greater understanding about the situation and improve the accuracy of generated annotations.

The research in this chapter regarding new technologies and algorithms provides novel ways in which a remote collaboration system can be implemented. Each technology has its benefits which can be leveraged to meet our requirements. Based on the investigation thus far, several mock prototype systems were conceived as shown in Figure 12. These prototypes focused on different modes for the delivery and input of information using HMDs, projectors, RGB and depth cameras mounted in various configurations.

The chest-mounted camera system, shown in Figure 12a, removes as much hardware from the technician's head as possible to increase his peripheral vision which is important in a dangerous work environment. Inversely, a head-mounted system, as shown in Figure 12b, frees the technician's hands, aiding in their ability to perform repair work. The off-body mounted system, shown in 12c, attempts to remove as much restriction on the technician as possible at the expense of the technician's ability to readily access complex annotation information and the situational awareness for the remote expert. Finally, a system could use an amalgamation of different hardware in a variety of different configurations to maximize their effectiveness.

(a) Chest-mounted camera and hand-held displays for input and display of information.

(b) HMD and camera with input being provided by gesture recognition.

(c) Off-body mounted system with hardware attached to a freestanding tripod.

(d) Amalgamated system utilising different elements from previous mock prototypes

Figure 12: Different modes for information gathering and display for the mock prototype system for the on-site technician.

First, Chapter 3 will discusses the work done to develop a framework on which the remote collaboration systems will be built. This work is common to both the systems that were developed.

3

FRAMEWORK

The remote collaboration system will consist of two units, an on- and off-site unit. The on-site unit is responsible for displaying virtual information to the technician at the work site and relaying information about the work environment to the remote expert. The off-site unit provides a view of the environment and provides a means for input for annotation data. Figure 13 shows the flow of information between the two units.

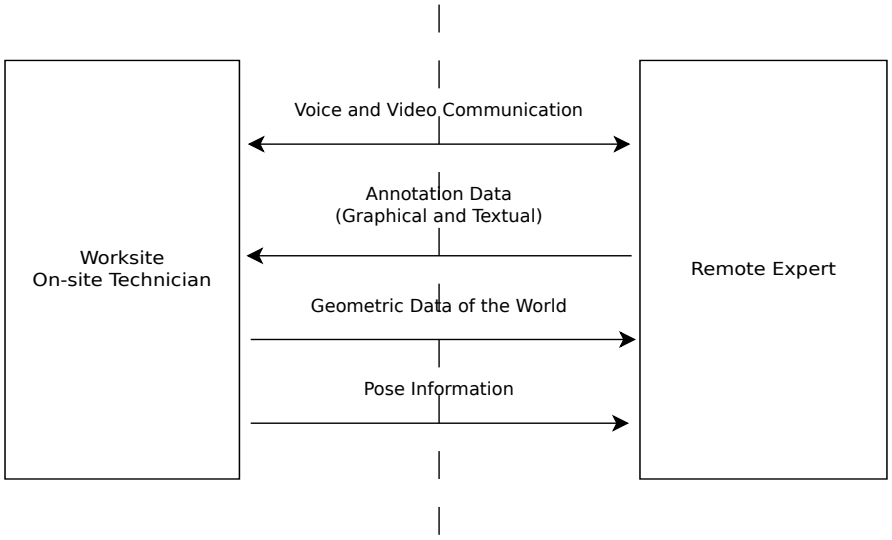


Figure 13: The generalized flow of information between the on- and off-site units. The left side of the dotted line represents the on-site location with the right side showing the off-site.

The hardware that is used to display the annotations for the on-site technician can be independent of the system developed. During the development process, two separate sets of hardware were developed to display the annotations. The details and development of the hardware for each separate system are discussed in their respective chapters, 4 and 5.

In this chapter, work on the application layer, which includes the communications system and the Graphical User Interface (GUI) used for the off-site unit, is detailed. After a review of the HITLab NZ system, it was found that the presence of lag greatly reduced the usability of the live annotations. The following sections will discuss the use of different video capture and streaming techniques as well as communications protocols to minimize these delays. First, however, guidelines for the User Interface (UI) are discussed.

3.1 UI CONSIDERATIONS

The UI for the remote user is focused on improving the experience for the off-site expert. The main goal for this development is to improve the situational awareness of the remote expert. The secondary requirement is the ability for the off-site users to create and control the display of annotations for the on-site technician, who will predominantly have their hands busy and have limited user input to the system. All of this information is displayed on the desktop PC that the remote expert is sitting at.

There are two main focuses of the UI. The first focus is on providing a view of the on-site environment. This can be in one of two different modes, a live-view or a map-view. The live-view provides a live camera feed of the work site from either a head mounted camera or a separately mounted camera that is off the body of the technician. This gives the most up-to-the-minute information about the environment. The alternate map-view provides a view of the environment that can be navigated independently of the camera. This ability is a vital requirement for increasing the situational awareness of the remote expert. To achieve this, extra 3D information from the hardware discussed in Section 2.2 and SLAM is leveraged to build a map of the environment. Because the type of display and capture hardware dictates how the map is built and navigated, each map-view is discussed in their respective Sections 4.3.2 and5.4.

The second focus of the UI is used to provide the input of annotations to be displayed for the on-site technician to see. Along with voice communication, this is the primary way in which the remote expert communicates with the technician. Annotations can be added when using either the live-view or the map-view of the environment. A mock example of the UI is shown in Figure 14 where the two focuses are shown in on the same screen. The UI for each system is different and shown in their respective section.

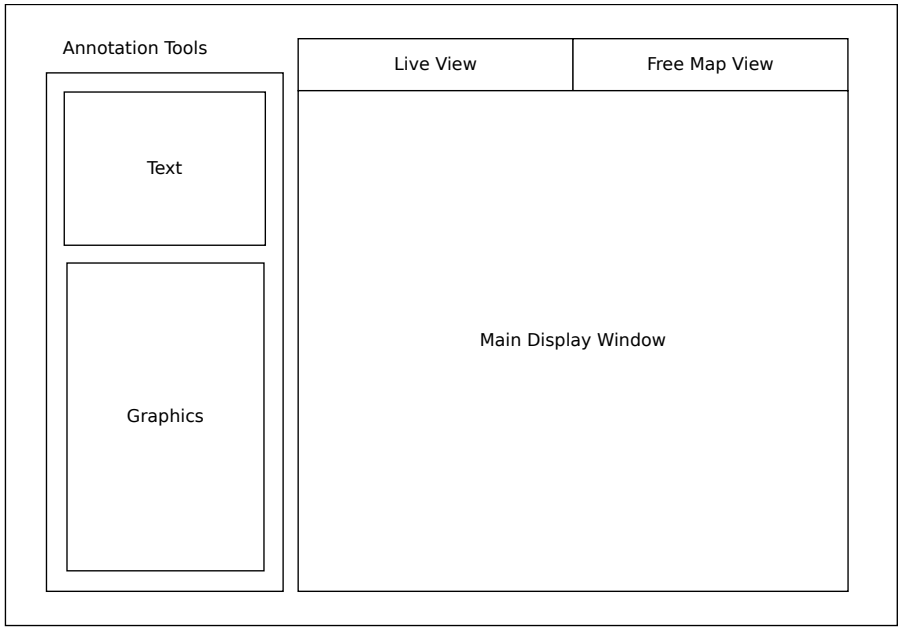


Figure 14: The mock UI which puts the main focus of the live and reconstructed view of the scene on the right. Annotation tools which can be used on either view is shown on the left.

3.2 COMMUNICATIONS

There are three main modes of communication between the on- and off-site users: live video, voice, and annotations. Because the task space for both users is the environment of the on-site technician, the flow of information for both the map data of the environment and the annotations is from the on- to off-site unit as shown in Figure 13.

Because annotations are generated by the remote expert and the hands of the technician are generally occupied with the task in front of them,

the onus is then on the remote expert to control the way annotations are shown. To do this in a way that is non-obtrusive to the on-site technician, annotations are, by default, set to have a limited viewing life. This means that after a small preset time, the annotations disappear. This reduces the clutter in the on-site technician's view.

The implementation of packet level communications protocol used to transport this data is discussed in the next section. The transmission of live streaming video is discussed in Section 3.2.2.

3.2.1 *Packet Messages*

A TCP connection is established from the on-site unit to the off-site computer which acts as a server. All information is serialized and sent using the Boost library [33, 3] and reconstructed on the other side of the communication channel. Inside the header which precedes each message, is a length parameter defining the size of the incoming packet. This gives us a maximum of ~4 Mb for payload data, which is ample for map and frame data. After each message is completely received, or a time-out is experienced, the receiving party then sends an ACK or FAIL message and waits for another header message.

To handle multiple message types, the start of the message is encoded with an enumeration of the type of data. The rest of the message consists of optional fields depending on the type of message that is sent. This is shown in Table 3.

Table 3: Bit allocations for Request Packet

INFORMATION	BITS
Header start	(0xAA0)
Length Header	22
Data Type	4
Payload	...

The party which initializes communication is responsible for en-queuing data and handling the guarantee of packet delivery. Data to be sent is prioritized in classes with an expiry condition attached to each data structure. If the expiry conditions are met while the data is still in the queue, it is discarded. This is particularity true for video frame data, which is discussed in the next section.

3.2.2 *Streaming Video*

Several methods were investigated for live video transmission. One method involved using a FFMpeg and FFServer to create a network socket that is available from the client machines to stream the video. However, because the FFMpeg process puts a lock on the capture device, the broadcast interface must be used to access the video feed on the local machine. This presents a 1~2 second delay, which was unacceptable for a HUD or hand-held tablet implementation. To solve this problem, a M-JPEG compression along with the packet expiry condition was implemented. The packet expiry conditions stipulates that: 1. There is only one live frame in the out-going queue at one time. When the new frame arrives, the old frame is removed and the new frame takes the same position in the queue. 2. When multiple map update packets are in the queue, live frames are always injected behind the next map packet to be sent. This implementation removes the delay on the local video loop and allows the most up-to-date information to be transfered to the off-site user. In times of high bandwidth demand (when multiple packets of map data are queued to be sent) the frame rate for the live feed automatically drops, allowing the map data to trickle through. Figure 15 shows the system in action streaming M-JPEG frames at a 640×480 resolution and 3D map data with ~400 map points and ~20 key frames .

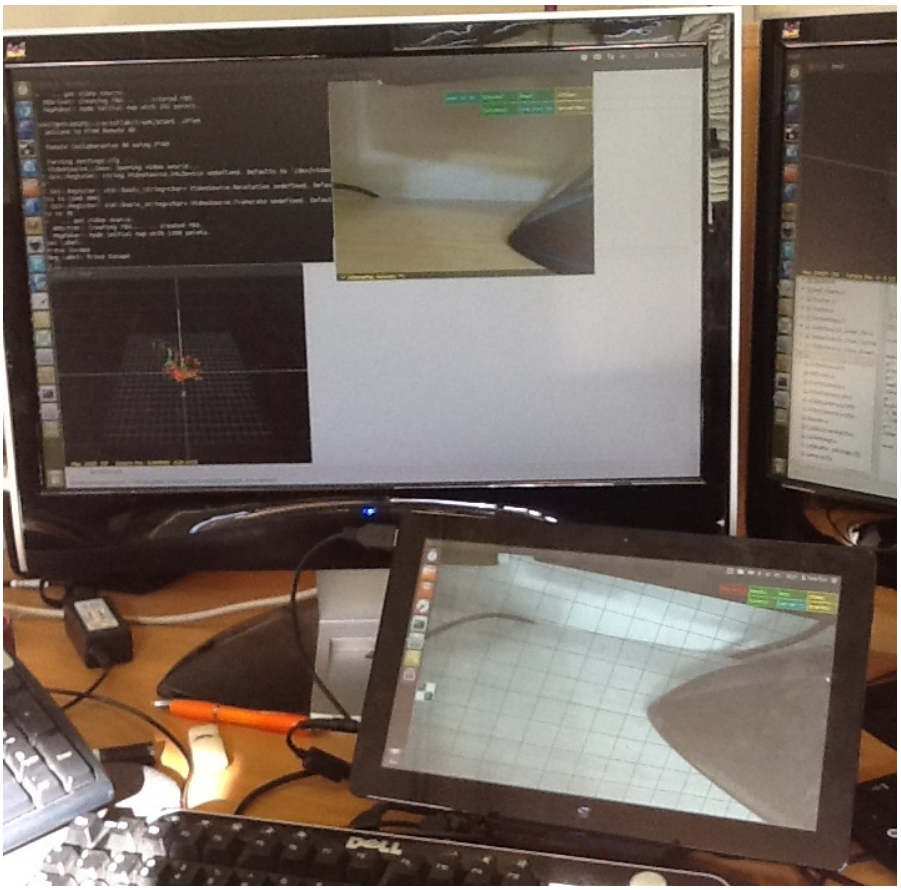


Figure 15: The video data is streamed over a network connection using the M-JPEG compression algorithm and priority enqueuing system.

4

HEAD-MOUNTED SYSTEM

Remote collaboration systems that use hand-held devices such as those developed by Platonov et al. [13] and the HITLab NZ group provide a very convenient and natural method for viewing and manipulating AR data; they also improve the awareness and peripheral vision of the on-site technician without the need for HMDs. However, these hand-held systems make it difficult for a technician to use both hands during complex repair tasks.

One of the main goals for the development of this first system is to incorporate a head-mount display and camera to free the use of the technician's hands while minimizing the reduction of the technician's peripheral awareness. Any hardware that is head-mounted must be optimized to reduce the impact on the user while still providing as much auxiliary information as possible. Other goals include developing a system that is highly mobile with lower power consumption to facilitate the use of the unit in the field. Finally, a robust SLAM algorithm must also be implemented to aid in the tracking of the technician's pose.

This chapter describes the work done to create the first system and fulfill the goals of a remote-collaboration system set out above. The process of hardware selection, implementation of tracking software, and development of a framework for communication and annotation will all be outlined in this chapter. A brief review is also included at the end of this chapter.

4.1 OVERVIEW

This remote collaboration system consists of two units, on-site and off-site, linked by a network connection. Besides the standard voice communication between the two parties, the on-site unit utilizes a HMD which allows the technician to view AR annotations while keeping their hands free to perform any required work. The implementation of the PTAM algorithm ensures that the annotations displayed are relevant to the view without the need for planted fiduciary points. The off-site user receives a live video feed of the on-site unit via a camera mounted to the technician's head, along with a reconstructed view of the remote environment built using a slightly modified version of the PTAM algorithm. Together these two pieces of information help to give the off-site user a better understanding of the work environment and increased situational awareness. Figure 16 shows the flow of information between the two users. The current capabilities of the system are also listed below:

ON-SITE The on-site client can use either a hand-held or heads-up display that will overlay virtual annotations onto the user's view of the environment. It is currently able to:

- Track and model the world with sparse 3D data. It is able to deal with with rapid movement and to regain tracking of the camera position when tracking is lost.
- Deal with a non-static environment which may feature background and foreground movements across the screen, including obstructions with the user's hands.
- Display simple annotations, including drawn graphics and text, as AR elements attached to the real world.

OFF-SITE The off-site server accepts connections from on-site units and will have a virtual re-creation of the on-site environment. It is currently able to:

- Reconstruct the scene with limited 3D detail textured with full resolution screen captures from the world.

- Completely control the annotations visible to the on-site user. This includes drawing up, sending, and removing annotations from the on-site user's environment.

- Virtually navigate the environment using previous frames of video, or the 3D reconstruction.

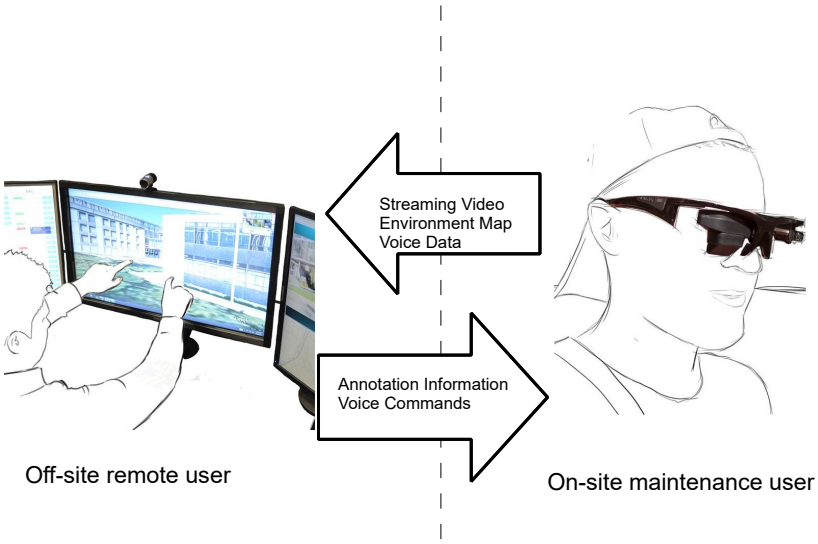


Figure 16: Overview of the operation of the system. The flow of information is shown with a solid arrow. The dotted line separates the on- and off-site locations.

4.1.1 *System Architecture*

The current system overview and the flow of data around the different modules are shown in Figure 17. The on-site unit uses the current pose estimation to draw annotations on the screen which are anchored in the real world. A collection of selective frames are sent back to the off-site unit to be used in the reconstruction of the on-site scene.

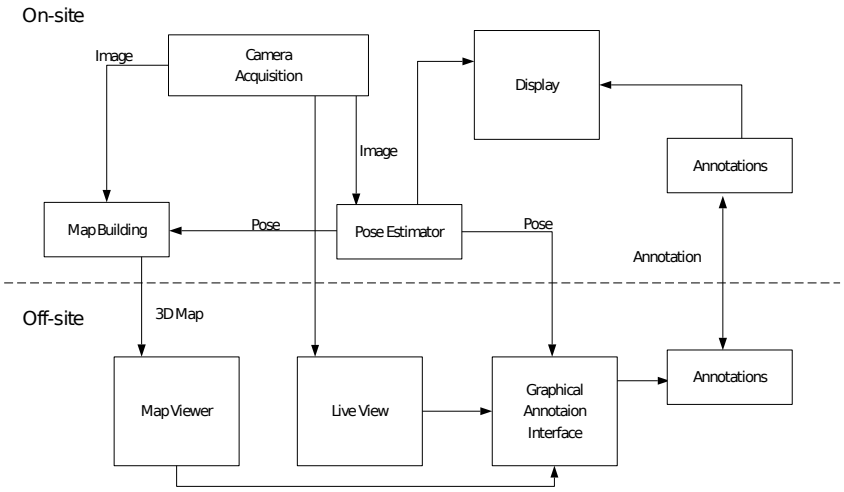


Figure 17: Overview of the system architecture for the head-mounted system. The flow of information is shown with a solid arrow. The dotted line represents the separation between the on- and off-site unit.

The following sections, 4.2 and 4.3, describe the client and server side implementations in more detail.

4.2 ON-SITE UNIT

This unit features a camera to gather information and a heads-up display in which annotation information is displayed as an overlay in the real world. A single RGB camera captures images at 30 fps with a resolution of 640×480 pixels. It is positioned to capture the user's view for localization and transmission to the off-site user.

4.2.1 *Calibration and Initialization*

For the camera to be used effectively in localization, sub-pixel calibration is required to produce usable information during pose calculation. Zhang's pin-hole camera model [41] was employed, which utilizes radial tangential distortion. Using perspective transformation, we can calculate the distortion of 3D points that are projected onto

an image plane. This can be represented with k_1, k_2, k_3 as the radial distortion coefficients p_1, p_2 as tangential distortion coefficients. A calibration matrix to compensate for this intrinsic distortion can be calculated from a known image. In this work, a chessboard pattern of known dimensions is used to calculate this. Information about this calibration is included in Section A.2.2 of the Appendix.

To calculate epipolar geometry of the world using only one camera, a translated view of the same environment is used to initialize the map as described by both Klein and Stewenius [18, 34]. Due to this, the same assumptions of the environment, having a general planar surface, applies. However, after initialization, the requirement for a general planar surface no longer applies. This means that thought the scene may not be planar, any small planar surface in the environment can be used for initialization, after which the algorithm functions without this assumption.

4.2.2 *Display Hardware*

In order to have the minimal amount of hardware attached to the technician's head, work was done to optimized a HMD. A Vuzix Smart Glass[1] was modified to be a monocular display with a 3D printed casing as shown in Figure 18. An occluded display was chosen to ensure that alignment was achieved between the annotations and the view of the world. A feat which can not be robustly achieved with see-through displays due to the inability to detect the position of the display relative to the user's eye. The modified monocular Vuzix display also offers the technician much more peripheral information compared with stereoscopic type displays which covered both eyes. The display covers the dominant eye and offers a view of the world, captured by the RGB camera, with small annotations overlaid on top. By displaying the annotation view of the world to the dominant eye, the brain is more able to seamlessly merge that view with the view

1 The Vuzix Smart Glass is a Stereoscopic type Occluded display. The specifications were listed in Section 2.2.1.

of the real world. It should be noted though that around 30% of the population is left-eye dominant, causing them discomfort and the inability to focus when using right-eye displays [22].

Due to the limited size and resolution of the Vuzix display, a fully occluded stereoscopic type display manufactured by Sony was also used in this research for comparison. The Sony HMZ-T1[2] offers dual 720p displays for each eye with a 45° viewing agle. This display with a modified mount to carry the wide-angle 160° camera is shown in Figure 19.

Figure 18: The monocular viewing system. A Vuzix Smart Glass display is modified to remove one eye piece. It is replaced with a 3D printed housing to form a monocular opaque display.

A more detailed review of the use of the system and the different display hardware and feedback is given in Section 4.4 of this chapter.

4.2.3 *Tracking*

To overcome the problems in Section 2.1.1, where the targets for annotations are not fixed to the world but to the frame of the video, the

2 http://www.sony.com/product/hmz-t1

Figure 19: The Sony HMD viewing system which offers a reasonable FOV. The wide-angle camera is fixed on top to offer the user a view of the world.

SLAM algorithm called PTAM [18] was used. This tracking algorithm locates the position of the camera in the real world while simultaneously building a sparse map of the environment. When viewed on the HMD, this allows the annotations to be anchored in the real world, staying fixed on a location regardless of the camera position. This gives a better sense of immersion. The tracking algorithm can be seen in action in Figure 20.

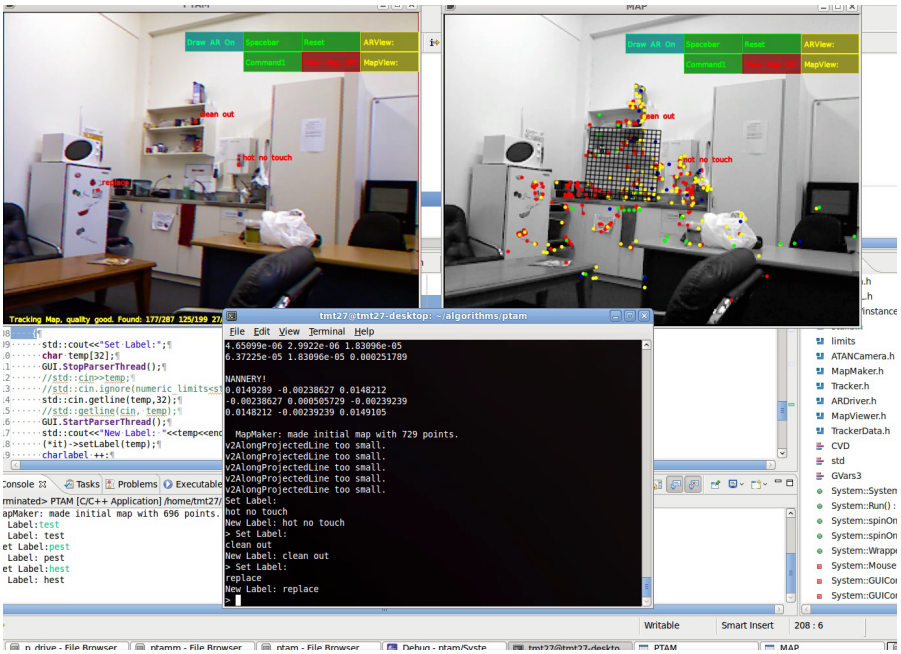


Figure 20: This figure shows the tracking and reconstruction algorithm in action. The top left frame shows an annotation anchored in the world. The top right shows natural planes for initialization found in a complex environment.

Though this sparse mapping algorithm provides less depth information about the environment, it was chosen to satisfy the requirements in Section 1.1 for two main reasons. First, it requires only a single RGB camera to operate, reducing the cost of the system. Secondly, it has very modest computational power requirements. The system can run on a small laptop, greatly increasing the mobility of the system. The advantages and disadvantages of this algorithm are listed below:

Advantages

- This system is able to operate as a head-mounted system due to the active camera tracking.
- It is able to operate without any previous knowledge of the world or artificial fiduciary points.
- It can be readily deployed with minimal extra cost for the maintenance engineer.

DISADVANTAGES

- Tracking of the world can sometimes be lost, which makes the system ineffective and disorientating while it tries to regain tracking.

- The system only works on the premise of a general planar surface.

While the PTAM algorithm operated effectively, it occasionally lost tracking. When this occurred, tracking could be regained in most cases when the camera crossed over the pose of a previous key-frames. However, because the on-site user does not have access to the map that shows the location of previous key-frames, it sometimes takes a few moments of moving the camera around aimlessly before tracking is regained. Because of this, work was also done to try and improve tracking with the use of IMUs. However, since this was not successful, interim results of this work is included in Section A.1 of the Appendix.

4.2.4 *Threading Detail*

As shown in Figure 21, the operation of the client software is separated into three main threads: Tracker, MapMaker, and Communications. This allows the system to run in real-time while dealing with any system delays with minimal impact on the operation of the entire unit. The Tracker thread is responsible for gathering new image data, estimating the current pose of the camera and drawing annotations to the screen. When tracking is lost, the tracking thread is also responsible for re-localization. The MapMaker thread takes any new frames and builds them into the map of the world when required. Similar work is also described by Gauglitz et al. [8]. The communication thread is responsible for relaying information between the client and the server. The timing for each process is shown in Table 4, which allows for a 33 fps operation.

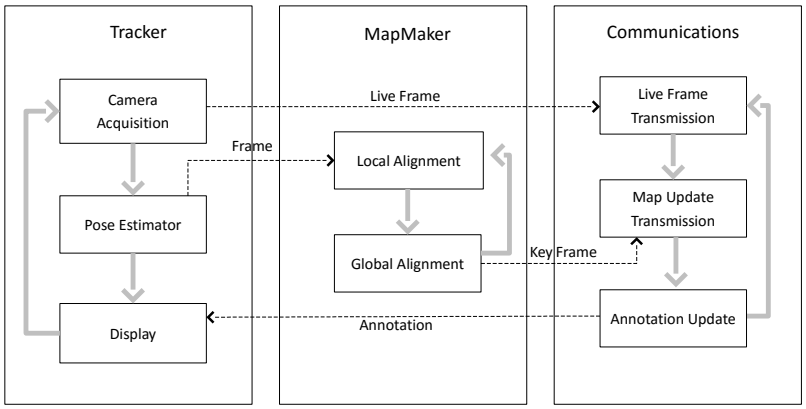


Figure 21: Main processing threads of the on-site system. The operations of each thread is detailed in the box. The dotted line shows the communications between the different threads.

Table 4: Average processes times for operations

ACTION	TIME(MS)
Frame to Frame (Tracker)	30
Local bundler Adjustment (MapMaker)	270
Global bundler Adjustment (MapMaker)	1700
Keyframe Re-localization (Tracker)	20

New key-frames created by the original PTAM algorithm are also used to update the reconstructed environment for the off-site user, which is discussed in more detail in Section 4.3.1.

4.3 OFF-SITE UNIT

The off-site unit acts as a server to accept connections from on-site clients. It shows the client's Point of View (POV) as well as a more complete view of the environment to increase the situational awareness of the remote expert. The overview of its functionality is primarily covered in Section 3.1. The system is also responsible for the full 3D reconstruction of the environment based on the data received from the on-site unit. In the following subsection the 3D reconstruction algorithm is discussed.

4.3.1 *3D Reconstruction*

Situational awareness for the off-site remote expert can be greatly improved by providing a representation of the environment that is not limited by the POV of the on-site technician. This is done by interpolating the sparse 3D information of the world obtained from the PTAM algorithm. By developing this algorithm to generate 3D maps rather than using extra cameras to calculate epipolar geometry as done by Poelman et al. 2012 [31], the amount of hardware that is mounted to the technician's head is reduced. This reduction in bulk and weight can improve the comfort of using the system.

The mesh for the 3D map is built using 2D Delaunay Triangulation algorithm from the CGAL library [6]. To create a mesh in $\mathbb{R}^3$ using the 2D Delaunay algorithm, triangulation is first calculated based on the view from the frame's pose location. The vertices are then projected back into $\mathbb{R}^3$. An example of this is shown in Figure 22. Section A.3

r29 of the Appendix shows the code for 2D triangulation which is performed for each key frame.

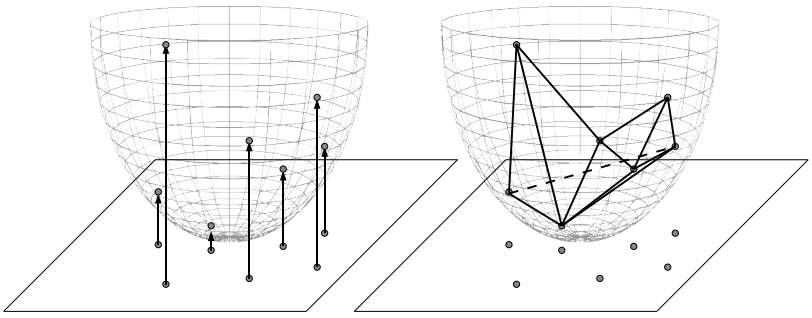


Figure 22: Projection from $\mathbb{R}^2$ to $\mathbb{R}^3$ using natural feature points from the environment [42]. x and y dimensions are in the plane at the bottom with the z dimension rising up.

As each new key frame is added to the model, a mesh is created that covers new feature points. An example of a sparse 3D reconstruction from a single key frame is shown in Figure 23. The pseudo code listings are shown in Listings 1, 2 and 3.

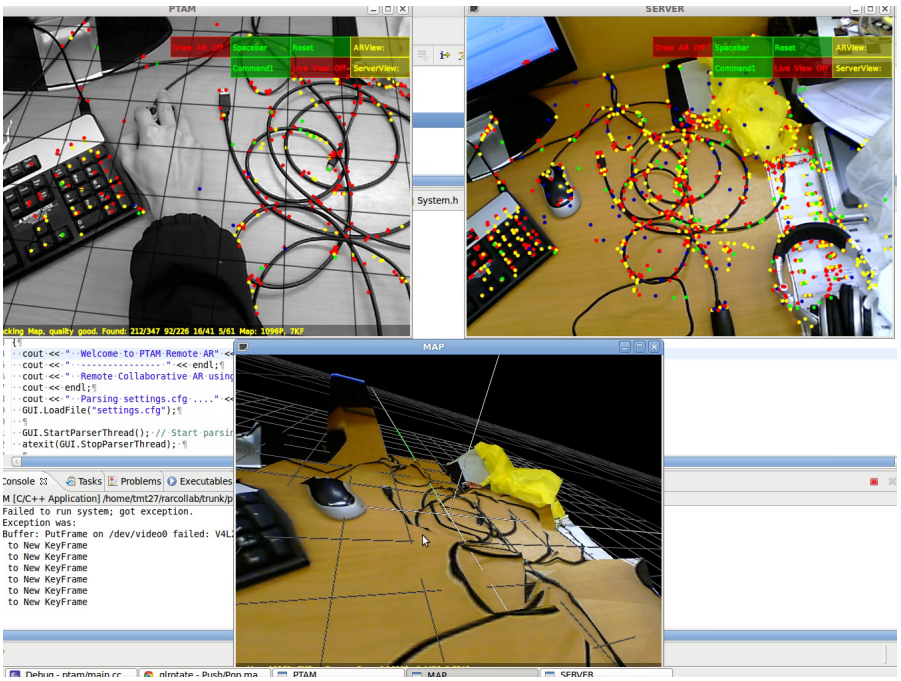


Figure 23: 3D reconstruction of a single frame using natural feature points from the environment.

Listing 1: Triangulation of a new frame from the camera

```
triangulate(new_frame){
   if (faces is empty){
      create face at random
   }
   for_each (vertex in new_frame){
      add_vertex(vertex)
   }
   for_each (face){
      if (face is not textured)
         add texture from new_frame
   }
}
```

Listing 2: Addition of a new vertex into the current mesh

```
add_vertex(vertex){
   for (each face){
      if (vertex is inside face){
         split face into three
         split original texture
      }
   }

   %face is outside of convex hull
   restore triangulation by flips
}
```

Listing 3: Removal of a new vertex from the current mesh

```
remove_vertex(vertex){
        for (each face){
                if (vertex is inside face){
                        remove vertex in triangulation
                        re-triangulate hole
                }
        }

}
```

Errors in the mesh, as seen in Figure 23, are usually seen as sharp spikes that fall toward or away from the camera in the z axis. As the map is built up and refined, erroneous feature points are automatically removed by the PTAM algorithm. When these points lie within a mesh they are automatically removed, flattening out the overall surface.

As multiple key frames are generated, they are stitched together as shown in Figure 24. Due to the operation of the 3D reconstruction algorithm, the map generated is biased towards building the smallest sphere centered around the camera where no 3D information is available. This also means that annotations that are anchored between natural feature points tend to be inaccurate in the z direction.

4.3.2 *Map Display and Navigation*

The main goal of the hardware on the server side is to accurately reproduce the on-site environment to allow the remote expert to assess the situation and direct the on-site technician. As shown in Figure 25, the two main types of information that are sent back to the off-site unit are a stream of the technician's POV and the 3D reconstruction of the environment.

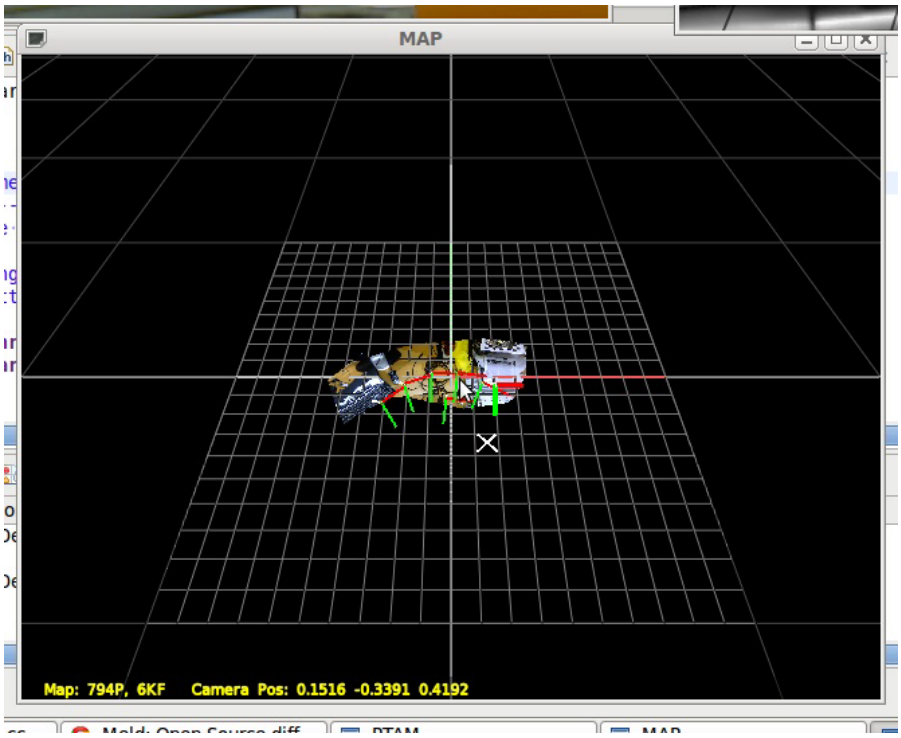


Figure 24: 3D reconstruction of a scene using multiple overlapping frames. Key-frame locations are shown using the RGB orthogonal axis.

The map view can be navigated in two different ways. The first way is via the 3D reconstruction of the world as shown in Figure 24. The user is free to move the position of the camera using the mouse cursor and different modifiers such as Alt. and Ctrl. for zooming and panning respectively. The second way is by jumping between key-frames using the arrow keys. This jumps between viewpoints that were constructed when a key-frame is taken and hides inaccuracies of the z calculations.

New annotations are added to the map simply by clicking in the right view window. Because of the sparse density of 3D points on the map, annotations are anchored in the world based on a weighted average of the three closest points as shown below. $\mathbb{R}^3_{(x,y)}$ is taken directly from the perpendicular plane and $\mathbb{R}^3_{(z)}$ is defined as:

Figure 25: Differing view of the scene given to the off-site and on-site user. Clockwise from the top-right, the image shows the view of the on-site user, the view of the off-site user, a navigational map of the entire scene showing key-frames and natural markers.

$$\mathbb{R}^3_{(z)} = \frac{\sum_{i=1}^{n} \left(\frac{z_i}{d_i} \right)}{\sum_{i=1}^{n} \left(\frac{1}{d_i} \right)} \tag{1}$$

where

- n is the number of neighboring points to consider
- z_i is z value of the point i
- d_i is the distance to the point in the (x, y) plane

4.4 REVIEW

A remote collaboration system using HMDs was successfully created. The on-site user views annotations left by a remote expert on a pair of head-mounted stereoscopic displays. Their view of the world is supplied by a 170° wide-angle camera. Annotations supplied by the remote expert are anchored in the real world using interpolation from the sparse 3D data supplied by the PTAM algorithm. The off-site expert receives a live view of the environment along with a 3D reconstruction of the environment. The expert can navigate the view of the environment independently of the view of the on-site technician.

The use of the hands-free HMD was very beneficial to the on-site technician in terms of executing the work, though the lack of ability for the technician to manipulate annotations was somewhat limiting. This meant that the onus was on the off-site user to keep the annotations generated non-obtrusive. The on-site technician then relied on the live video and audio feed to communicate any other information.

The ability to anchor annotations to fixed points in the world proves to be much better for conveying complex repair tasks and information to the on-site technician. However, this benefit was negated when the SLAM algorithm occasionally lost tracking. This was more likely to occur during fast movements of the camera, bad map initializations and significant changes in the work environment.

Limitations such as the requirement for a planar surface during map initializations did not limit the system too much. In most environments, a planar surface, even one that was relatively small, could always be found and used to initialize the map. After initialization, the PTAM algorithm had no problems dealing with complex environments. Bad initializations did occasionally occur and compromised the stability of the pose tracking but could easily be detected by jittery tracking and the map could be readily reinitialized.

The use of interpolation on sparse 3D data from the PTAM algorithm to build mesh maps was novel work. This reduced the computational cost and communication load for running the system. It also allowed utilization of common off-the-shelf RGB cameras instead of more expensive depth cameras. However, this method for generating 3D data of the environment also meant that more pronounced errors were generated in the map compared to those generated by depth cameras.

Though the sparse mapping did mean that there was limited definition and sporadic errors in the z direction of the map, their effect on the user's understanding of the environment was negligible. There are two reasons for the limited impact of sparse mapping. The first reason is that other visual cues in the texture of the mesh helped to determine shape. Secondly, because the off-site user spent large portions of their time viewing the environment at poses similar to the original key-frame, this similarly mitigated the perceived displacement error resulting from incorrect z values.

The ability for an off-site expert to independently navigate a map of the environment was very helpful in terms of conveying situational awareness. However, because the 3D mesh map was updated only when new key-frames were generated, the map could easily become stale in a dynamic environment. This forced the off-site user to spend more time in the live-view to get information and use the map-view to draw and supply information.

The live-view from the on-site user's point of view was slightly disorientating for a remote expert to use. This happened despite the fact that a wide angle camera was used and that it was mounted on the technician's head. Even with a person's natural instinct to stabilize their own head, any conscious movements of the head proved to be disorientating for the remote expert.

The current selection of off-the-shelf HMDs proved to be slightly limiting for the on-site user to use. First, the size and weight[3] of the HMD tended to be encumbering for the user. Secondly, the unnatural 45° FOV of the HMD and camera also caused slight disorientation, especially when used for long periods of time. The monocular Vuzix display also featured similar discomfort with prolonged use. This was also accentuated by the fact that both eyes received a view of the world with different FOVs. The limited size and resolution of the display also reduced the usefulness of this display. This ultimately meant that complex tasks were harder to perform as concentration was divided and situational awareness was still limited.

A disconnect of the remote expert's "presence" could also be felt. Users would feel that the expert was merely a "voice in your head" and annotations would just pop up in the world.

The limitations that were discovered in this system were the catalyst for the development of the next projection-based system described in Chapter 5. This system attempted to remove the issues involving the use of HMDs, head-mounded cameras, and disconnected feeling of "presence".

3 The HMZ-T1 has dimensions of $210 \times 196 \times 110$ mm and a weight of 420 g without the processing unit.

5

PROJECTION SYSTEM

The use of a HMD discussed in the previous chapter decreased the situational awareness and ability to focus on a complex task for the on-site technician. This is because it replaces the view of the dominant eye or both eyes with a lower resolution, and smaller FOV image. The view from this HMD's camera could also become disorientating for the remote expert to use. In this chapter an alternative solution which uses projection technology is investigated.

The system discussed in the previous chapter utilized a head-mounted display to show the information to the maintenance engineer. An alternative method of displaying annotations would involve the use of projectors. This alternative system would require no extra equipment on the maintenance engineer's body but simply projects the AR information directly onto the real world using over-the-shoulder laser projectors for the technician to see.

After initial investigation, two papers were found to have done similar work [29, 9] with regards to displaying AR information. My system extends their work by incorporating 3D model data to allow much more accurate projections on a larger range of more complex surfaces. This allows for the system to be used in a wider range of scenarios present in an industrial setting, which is likely to feature irregular surfaces.

The system developed in this chapter, while auxiliary to the scope of development for the project specified by the HITLab NZ, would make a very compelling concept system for future technologies. It should be noted that a class of systems that utilize projection technol-

ogy to display information onto complex non-planar surfaces also go under the name of projection mapping. Currently, the use Projection Mapping is predominantly used in art displays and exhibitions [14].

We attempt to utilize projectors to display augmented information in the real world to aid in a more practical task of remote collaboration. This chapter details the work done to use projection mapping in our augmented reality remote collaboration system. An overview of the system is given in Section 5.1. This is followed by the hardware specifications. The work on reconstruction of the scene was done in conjunction with Mathew Tait, another member of the HITLab NZ group, and discussed in Section 5.4. Finally, a review of the system is conducted in Section 5.6.

5.1 OVERVIEW

The projection system can link the on-site technician and the off-site expert via the same audio and video link as described in the previous chapter. The main difference is the way in which the annotation information is displayed to the on-site user who now, instead of a HMD, has a tripod set up behind them in an over-the-shoulder position. On top of the tripod is mounted a depth camera and projector on a 2 axis gimbal. This can be seen in a mock up of the system in Figure 26 and discussed in more detail in Section 5.2.

The gimbal-mounted pod acts as the "eyes" of the remote expert and as an extension of their arms for pointing and gesturing. The information supplied to the remote expert is a live video feed and a dense 3D reconstruction of the world. The directions are then displayed back to the technician on site as drawn annotations, which are projected directly onto the real world using the on-board projector.

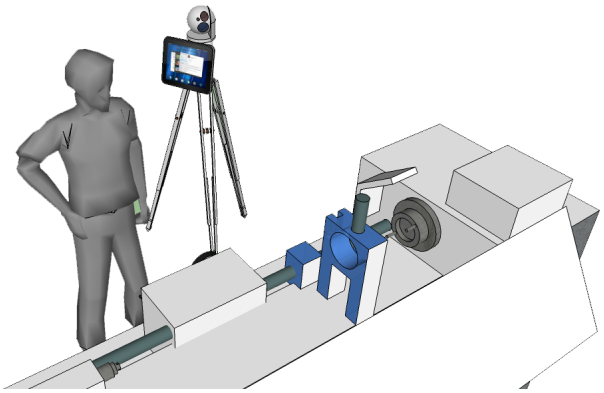


Figure 26: A prototype example of the laser system with the camera and projector mounted on a 2 axis gimbal.

The current capabilities of the system are also listed below:

ON-SITE The on-site unit features a projector and Kinect mounted on top of a 2-axes gimbal. It is currently able to:

- Track and model the world with dense 3D data. It is able to deal with large scene changes and obstructions to the view.
- Annotations can be accurately displayed on almost any surface of the environment. The use of depth data from the Kinect camera allows us to compensate for complex geometries.
- Both the projector and the Kinect camera have extended FOVs beyond their native range due to the use of a 2-axes gimbal.

OFF-SITE The off-site unit re-creates the on-site environment for the remote expert, on which they are able to create annotations. It is currently able to:

- Reconstruct the scene with a dense 3D detail and texture it with full resolution frame captures of the world.

- Completely control the annotations visible to the on-site user as well as control the movement of the 2-axes gimbal.
- Virtually navigate the environment independent of the placement of the tripod.

5.1.1 *System Architecture*

The architecture of the Kinect-projector system also shares many similarities to the HMD system discussed in Chapter 4. However, the first main difference is the use of a depth sensing Kinect[1] camera which is used instead of the RGB camera. The second difference is the SLAM algorithm which is now implemented using KinFu [25]. This uses an Iterative Closest Point (ICP) algorithm on each pixel of the depth camera to give a dense 3D map of the environment. The third difference to the HMD system is the mounting of all the hardware on a tripod and the inclusion of a gimbal which controls the movement of the camera and projector and can be accessed by either the on- or off-site user. An overview of the system architecture is shown in Figure 27.

5.2 HARDWARE

Limitations of current projection technologies were the chief deciding factor during the selection of components and implementation of the system. Above all, the system must be viewable in a brightly lit room and after the initial investigation of projection technologies, the use of laser galvanometer would be ideal. However, due to constraints of development time, we opted to use a DPR projector as shown in Figure 28.

1 http://dev.windows.com/en-us/kinect

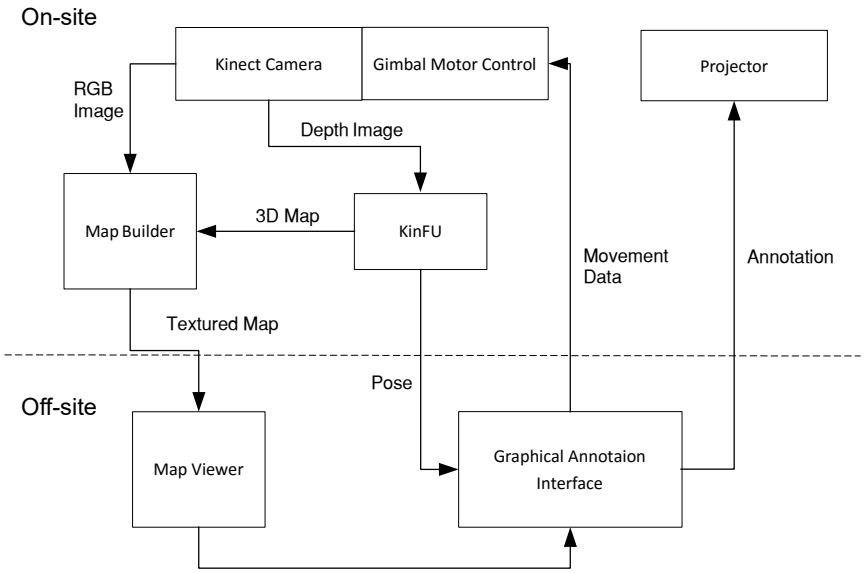


Figure 27: System overview of the projector system. The dotted line represents the separation between the off and on-site unit.

The projector chosen for use was built by InFocus, model IN1126[2]. The specifications are as follows:

- 68×68 cm projection size at 2 m
- 200 lumens
- 0.3 Kg

The maximum graphics size is dependent on the distance from the tripod to the surface used for projection. Short throw projectors can have throw ratios of up to $0.6 : 1$ and the Kinect camera has a throw ratio of $1 : 1$ as shown in Table 5.

Table 5: Throw ratios and projection sizes for the hardware positioned 630 mm away from the wall.

Component (Throw Ratio)	Horizonal X(mm)	Vertical Y(mm)
Kinect $(1 : 1)$	800	500
Projector $(0.6 : 1)$	320	200

2 http://www.infocus.com/projectors/ultra-portable/IN1120-Series/IN1126

Figure 28: The level of contrast achieved with the projector in a brightly lit room. The annotation projected is a red line in the center of the picture to the left of the bookcase. While the annotations were visible, the lack of contrast limited the complexity of annotations that could be viewed.

To overcome the drawback of the limited throw ratios of the projector and the Kinect camera, a gimbal and motors were used. Another option to overcome the limited throw ratios would be to use concave optics in front of both the Kinect camera and the projector. This would, however, proportionally reduce the resolution of both the captured images and the projected annotations. Therefor the option of fixing the sensors on a gimbal was chosen.

Robotis Dynamixel DX-117[3] servo motors controlled via a FT232R[4] FTDI USB to serial UART interface chip were used to control the direction movements of the camera and projector. This offers 300° of freedom in the horizontal direction and −40° to 60° freedom in the vertical direction on top of the FOV for the sensors. Hard limits for the movement of the servos are implemented in software. These prevent pinching hazards for the on-site user as the gimbals can be controlled by the off-site user. The different prototypes can be seen in the following Figures 29 and 30. More detailed information about the hardware and gimbals used is included in Section A.4 of the Appendix.

3 http://support.robotis.com/en/product/dynamixel/dx_series/dx-117.htm
4 www.ftdichip.com/Products/FT232R.htm

Figure 29: First prototype for the Kinect-projector system. The system uses a laser pico-projector mounted on top of the Kinect camera.

Figure 30: The second prototype for the Kinect-projector system. The system uses an Asus XTION PRO LIVE depth camera which has reduced power consumption compared with the Kinect camera and a 200 lumen micro-projector.

5.3 CALIBRATION

To obtain the intrinsic properties of the projector along with the rotational and translation matrix between the projector and the camera, the calibration technique from Cámera Lúcida was used [30]. This data is then used to compute the view-port transformation required to display the annotations. More details on how the calibration data is used to display information are discussed in the next section. Calibration values for the system are included in Section A.2.1 of the Appendix.

5.4 RECONSTRUCTION

The work on the reconstruction of the scene was done in collaboration with Mathew Tait, another member of the HITLab NZ group. The on-site environment is reconstructed using the KinectFusion algorithm developed by Newcombe et al. [25], which gives a dense 3D reconstruction of the scene. This algorithm was chosen because it was easily implemented and it shortened our development time significantly. However, its heavy reliance on a GPU for computation makes it a poor choice for a mobile application.

Using the 3D mesh data from the KinectFusion algorithm and color data from the RGB camera, we can reconstruct a fully textured scene of the on-site environment. However, because of the offset of the IR camera and the IR projector, shadows where there is no information from the scene are inevitable. Objects closer to the camera cast two separate shadows where different types or information, either depth or RGB, are unavailable. This effect is shown in more detail in Figure 31. The resulting effect of these shadows are shown in the reconstruction in Figure 32.

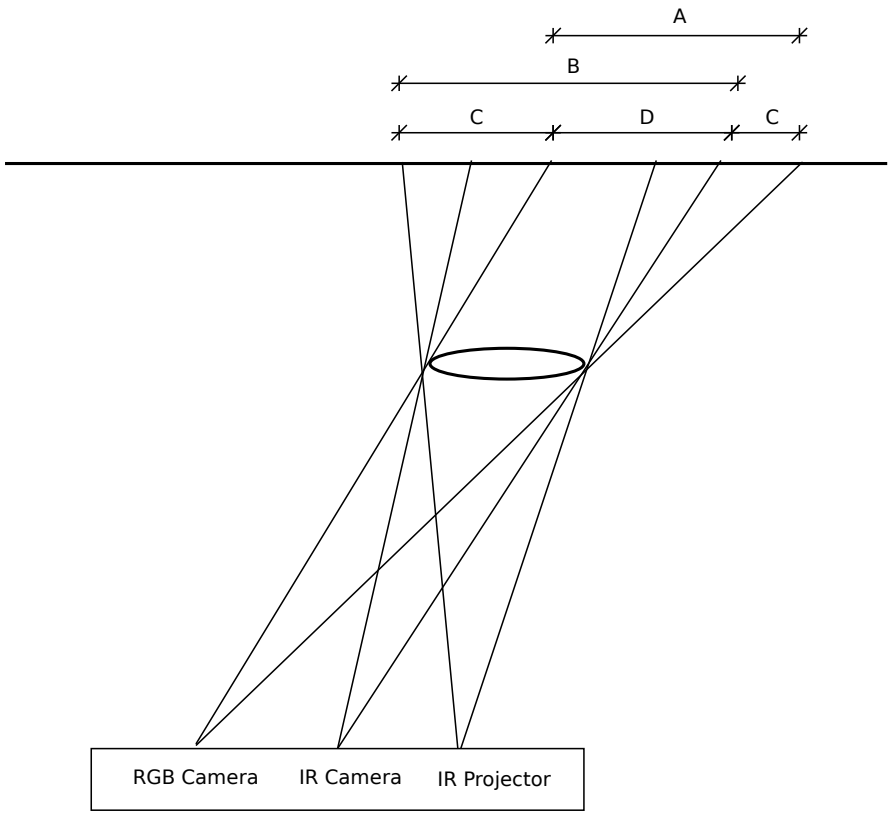


Figure 31: The Kinect's shadow effect shows how a near object can create wider shadow and how the offset of RGB and depth cameras create misalignment with the different types of data from the depth camera. Section A: missing RGB data; B: missing depth data; C: RGB data but no depth data; D: no data of any kind.

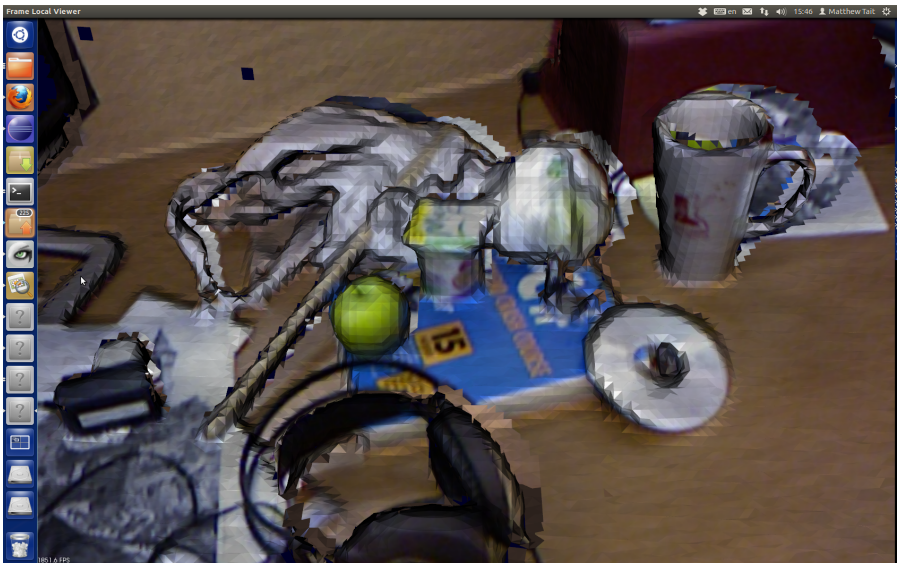


Figure 32: Scene reconstruction using textured Kinect Mesh data. The effects of the shadows and alignment of RGB and depth can be seen clearly around the edge of the objects. This figure was courtesy of Mathew Tait.

5.5 DISPLAY

Using the dense 3D information from the scene, projections on non-perpendicular and irregular surfaces are automatically compensated for. Points that are drawn on the 3D reconstruction of the scene are simply redrawn using the rotation and translation matrices calculated during the calibration. An example of this is shown in Figure 33. The scene information with only the annotations is then sent to the projector, which in turn projects that information directly onto the world.

5.6 REVIEW

We investigated an alternative way to display annotations for our remote collaboration system. A micro-projector is used to project annotations directly onto the real world. The projector is mounted on a tripod with a depth camera with a 2-axis gimbal. This system has many different benefits and drawbacks in comparison with the HMD

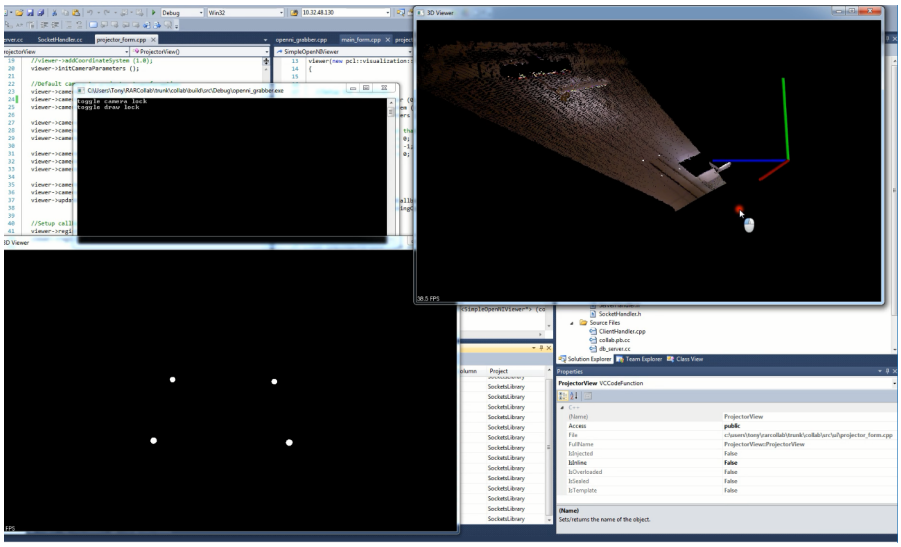


Figure 33: Points drawn on the 3D scene can then be directly projected back onto the real world. The top right window shows the 3D reconstruction of the world. The bottom left windows shows the information sent to the projector. Because the projector was mounted in an inverted position relative to the camera, the points closer to the projector were on the top of the projection display and thus spaced closer to compensate for the key-stone effect.

system. While this system was limited in terms of portability and visibility of annotations, it showed promise in terms of future development.

Using a projector to display the annotations offered several advantages over the use of HMDs. Firstly, because all of the hardware for the on-site unit is mounted on the tripod, the technician is not limited in any way in terms of their movement and their performance of the repair work. Secondly, the technician is not impeded by an unnatural FOV, cumbersome extra head-mounted equipment, and limited situational awareness of his surroundings because the annotations are projected and viewed directly. Another advantage of this system is that it is much less likely to lose tracking, compared to the PTAM system because of its fixed nature. Finally, this system allows an almost unlimited number of people to view the information together without requiring extra equipment.

However, there were a few limitations with this system. The fist limitation is the type of surfaces that can be projected onto. While there are no restrictions with a HMD as to what surfaces virtual annotations can be viewed on, projected annotations, can not be viewed on heavily textured surfaces, such as shagged carpet, and on highly reflective surfaces. The second limitation arises from Kinect's use of the spatially encoded infer-red light to acquire 3D data. Because of this, the system cannot be used in direct sunlight or places with significant infra-red light noise.

Another problem with the system was the type of projector used and its power requirements. It required ~300 Watts for all of the components of the system during operation. This meant that the system needed to be plugged into a wall socket when in use. Any attempt to power the system with batteries would either required a very large capacity battery, making the system non-portable, or suffer from severely limited operating time.

The micro-projector, while generating enough lumens (~200) to be viewable in a brightly lit room, was still weak. It did not provide a great deal of contrast and thus limited the types of annotations that were clearly visible. Simple annotations such as bold lines and boxes were viewable but texts, unless they were largely sized, were strenuous to read.

The use of different display hardware provides a remarkably different experience for the on-site technician. The removal of hardware fixed to their body makes them feel substantially more unrestricted both in movement and ability to observe their environment. The placement of all the on-site hardware on a tripod also gives a better feeling of presence that a remote expert is there. The sensation is that of a person over your shoulder helping you and being able to point to locations in the environment. Situations, though, would arise when the technician will block the line-of-sight from the remote expert. This leads to the interruption of work when the remote expert asks for their point-of-view (tripod) to be moved, but adds to the sensation that someone

is helping you over your shoulder. The reduced likelihood that tracking would be lost, due to the static nature of the unit, also helps to reduce the distraction and increase the robust nature of operating the system.

The experience for the off-site user is also different in many ways. The inclusion of dense 3D mapping greatly improves quality of the mesh map. The use of the KinFu algorithm is also much more useful for the dynamic update of the environment. Any changes to the environment are automatically and gradually updated to the map due to the operation of the algorithm. This makes the map-view much more useful to the remote expert. The algorithm also automatically deals with movement of the tripod. However, during some instances, the live-view would occasionally be crossed by the technician, which can be temporarily disorientating for both the live-view and map-view.

6

SUMMATION

Two separate solutions for the problem of remote collaboration to assist with complex maintenance tasks were investigated and implemented. We focused on the use of components and algorithms that allowed the capture of 3D information to make a system that is aware of the geometry of the task-space it was operating in. Though, in terms of qualitative investigation, more could have been done, the application of new technologies, algorithms, and novel ways to apply these to a remote collaboration system is our main contribution.

Both systems shared a large portion of their functionality. A 3D map and streamed video of the on-site technician's environment were sent to an off-site remote expert in both systems. Both displayed annotations that were generated by the remote expert and sent to the technician in a way which is independent of the technician's view.

The final capabilities of the framework are listed below:

- The framework features no lag when viewing the video on the on-site unit.
- The video frames are sent to the off-site unit using M-JPEG compression and a custom frame drop algorithm.
- A custom communications protocol allows us to balance the bandwith between live-video data and other map data.
- Annotations can be added in either the, live-view or map-view of the scene.

- Annotations automatically expire to unclutter the technician's view.

The two systems differ predominantly in the way in which the on-site technician views virtual annotations. The head-mounted system uses a HMD to display annotations.The projection system uses a micro-projector and depth camera to project information directly on the world. Both systems implement a different SLAM algorithm to allow the on-site unit to be moved freely without the system losing track of its pose. The following two sections 6.1 and 6.2 will outline the results in development for each system including the benefits, drawbacks, costs and novel work done.

6.1 HEAD-MOUNTED SYSTEM

The use of the SLAM algorithm and 3D reconstruction greatly aided in the ability for users to remotely collaborate on a complex repair task. The PTAM algorithm tracked the pose of the head-mounted camera, and to generated a sparse depth map of the environment. Using this sparse depth information along with frames from the RBG camera, a 3D mesh of the environment was generated using Delaunay triangulation and $\mathbb{R}^2$ to $\mathbb{R}^3$ projection. The pose tracking allowed annotations to be anchored in the real world, making them to be more easily understood. The 3D mesh map of the world gave the remote expert much better situational awareness of the environment, especially compared to the live-view from the head-mounted camera.

The final capabilities of the prototype HMD system are listed below:

- The system can track and map the world in real with without the use of artificial fiduciary points.

- A novel algorithm uses the sparse depth data from the PTAM algorithm to generate a 3D map, reducing the need for extra hardware.

- The off-site user can navigate the scene, independent of the on-site technician's camera location.

- Head mounted hardware has been optimized reduce weight and increase situational awareness with the use of only a single RGB camera and customized HMD.

However, it was found that using a HMD to display the information was not fully conducive to a natural view of the world and could hampered the technician's ability to perform complex repairs. Though not done in this research, several suggestions can be made to improve the system. These are discussed in the Chapter 7, Future Work.

6.1.1 *System Costs*

The cost of the system has a significant impact on the feasibility of the system for widespread deployment in an industrial setting. The ability for parts of the system to be replaced at a reasonable cost for the sake of maintenance and upkeep is important for the viability of this solution. Table 6 below shows the total cost for the HMD system as well as a breakdown of hardware expenses. This can be compared to the projector system's cost which is listed in Section 6.2.1.

Table 6: Major components of the on-site HMD unit and their costs, as of August 2013.

Component	Description	Price(USD)
Samsung Slate	Tablet Computer	1,200
Sony's HMZ-T2	Head Mounted Display	1,000
Housing	Component Housing + Mechanics	300
Misc.	Sundries	100
Total		2,600

6.1.2 *Processing Times*

Table 7 shows the processing times for each part of the algorithm for a map with ~400 map points and ~20 key frames using our chosen hardware. Table 8 shows the further breakdown of the first item in Table 7.

Table 7: Processing times for operations

OPERATION	TIME(ms)
Frame to Frame	20
KeyFrame to Map	10
Local Bundle Adjustment	270
Global Bundle Adjustment	1,700
Keyframe Re-localization	20

Table 8: Breakdown of times for Frame-to-Frame Operation

OPERATION	TIME(ms)
KeyFrame preparation	2.20
Feature projection	3.5
Patch Search	9.8
Iterative Pose update	3.7
Total	19.2

6.2 PROJECTION SYSTEM

Based on the feedback from the implementation of the HMD system, the use of a projector to display annotations was investigated. A micro-projector and depth camera are mounted on a tripod with a 2-axis gimbal. A Kinect camera captures the technician's environment as a dense textured 3D mesh. The KinectFusion [25] (KinFU) algorithm is used to perform SLAM functionality. Finally, a projector is used to display the annotations directly onto geometrically complex surfaces in the real world using projection mapping, extending the work of Palmer et al. and Gurevich et al. [29, 9].

The final capabilities of the novel kinect-projector system are listed below:

- Annotations are projected directly onto the real world, removing the need for HMDs and allowing annotations to be view by an unlimited number people at the scene.
- The combination of the depth camera with the projector allows for more accurate projections. This compensates for key-stoning and projection onto complex 3D environments.
- The on-site scene is reconstructed with increased detail compared to the HMD system because of the use of the depth camera.
- The system has a extended field-of-view due to the custom built housing and 2-axes gimbal.
- The system is capable of dealing with major scene changes and obstructions of the camera by the technician without losing tracking.
- The location of the system is also automatically regained when the tripod is moved.

There are many benefits of using the projection mapping system for remote collaboration. The use of a projector is a more organic way of viewing annotations, removing the inherent discomforts of using a HMD. The projector also allows an almost unlimited number of people to view the annotations without extra hardware costs. There is a better sense of presence for the technician, and more situational awareness for the remote expert. The resulting system deals with dynamic scenes better due to the nature in which the implemented SLAM operates.

The ideas in this work involving the use of a depth camera in conjunction with a projector to aid in the task of remote collaboration in

an industrial setting was considered novel. ABB filed to patent this in 2014. The patent, titled: *Method and Video Communication Device for Transmitting Video to a Remote User* [38] was submitted in the USA, the EU, and the United Arab Emirates.

However, there several key improvements that can be made to further develop the system including the increase in view-ability and portability. These improvements are discusses in more detail in Chapter 7 along with lessons that were learnt during this project.

6.2.1 *System Costs*

The cost of this system, after an initial evaluation and selection of hardware, is listed in Table 9, which shows the approximate cost of the major components, as of August 2013. This can be compared to the HMD system's cost which is listed in Section 6.1.1.

Table 9: Major components of the projection system and their costs, as of August 2013.

Component	Description	Price(USD)
Asus XTION PRO LIVE	Depth Camera	200
DX117	µcontrolled Servos motors	400
InFocus IN1126	Pocket Projector	1,100
Micro-PC	Small form factor PC	1,200
Nvidia GTX 700	Performance GPU	300
Housing	Component Housing + Mechanics	500
Misc.	Sundries	100
Total		3,800

7

FUTURE WORK

Based on the review of the two remote collaboration systems, there are a few aspects which still need improvement. In this chapter, we will outline the suggested direction for the future development of this project as well as the lesson's learnt.

7.1 HMD SYSTEM

Due to the way the map building algorithm operates, the map-view of the world can quickly become out-of-date in a rapidly changing environment. The current map building algorithm does not consider the age or resolution of the key-frame when adding data to the map. Work must be done to update the map more intelligently when two textures overlap the same area. The selection would be based on the resolution and the time at which each texture was captured. The algorithm will select the texture with the higher resolution given that it was not captured too long ago. The weighted texture value W could be calculated by:

$$W = \frac{W_{res}}{d_z} + W_{time}\left(1 + \frac{N_{frm}}{N_{max}}\right) \tag{2}$$

where

W_{res}	is the resolution weight value
W_{time}	is the time weight value
d_z	is the z distance of the pose from the vertex
N_{frm}	is the elapsed number of frames between the new and original key-frame
N_{max}	is the comparator for elapsed frames

With the constant advancement of HMDs, the use of lighter displays with wider FOVs will also help to improve the experience for the technician. Matching the FOV of the camera and display could also help to reduce initial disorientation when using the system.

7.2 PROJECTION SYSTEM

The choice to use the KinFU algorithm was made predominantly to expedite the development process and had a few drawbacks. The first, was lag of 1~2 seconds when the gimbal was used to move the head and when the annotations realigning themselves. The second drawback was the power requirements of the GPU on which the algorithm operated. A better option would be to use a less computationally intensive ICP algorithm. Improvements could be made by using cylindrical regression during the calibration phase to calculate the offset of the gimbal's axes from the camera's focal point. This would help to remove the movement lag by giving a very good initial estimate point from which to start ICP calculations from. Improvements to the KinFU algorithm made by Kahler et al. [15] may solve the aforementioned problems. The algorithm, called InfiniTAM$^{\infty}$, optimizes the algorithm to allow it to be run at 40 fps on a NVIDIA K1-based[1] Android device.

Using the projector in combination with the RGB camera also provides an opportunity to develop a display algorithm which compensates for the color and brightness of the surface that is being pro-

1 http://www.nvidia.com/object/tegra-k1-processor.html

jecting onto. This would allow the projected annotations to be much more consistent across a wider range of surfaces.

The limited brightness of the projector led to an unsatisfactory contrast of annotations in a brightly lit room. Use of an alternative projector type would be recommended. A laser galvanometer could provide the contrast needed at the expense of annotation complexity as the galvanometer can only project vector graphics of a single color. Depending on the laser used, though, the power requirements of the system may still be too great to allow the system to runon battery power.

7.3 ALTERNATIVE APPLICATIONS

The hardware and system developed for this remote collaboration unit could also see use in a variety of different applications with a bit of modification.

First, the use of different hardware could significantly change how the system could be used. A pico-projector could allow the system to be shrunk to a hand-held flashlight-like device. It could also be mounted to a mobile platform rather than a tripod such as quadcopter or wheeled vehicle. Both these options would significantly improve mobility.

The system could then be used for a variety of remote collaboration applications beyond machine repair, such as crime scene investigation, or even urban search and rescue. The system could also be used without the collaboration aspect. Computerized manuals could automatically help a technician through a repair task by displaying relevant annotations, using image recognition to identify the state of a machine.

Finally, a user study should be conducted to quantify the qualities of the two systems beyond the preliminary feedback received. An

identical task should be carried out using each system. The ability to perform fine motor-control tasks as well as perform large scale actions under instruction from a remote user should be evaluated. A task with variable on-site conditions should also be presented to assess the remote user's ability to understand and evaluate the situation.

7.4 LESSONS LEARNT

The constraint of time during the development of a system was always a concern. There was always a trade-off between allocating more time to develop a better system and getting the project moving, and finding that balance was key. Developing the system in an iterative fashion made it much quicker to get the project going but also made it harder to maintain and expand in the future. The final system developed was somewhat monolithic and not very modular.

Calibration of the hardware is key to the performance of a SLAM algorithm. During the development, the complex lens construction of certain cameras meant that the distortion could not be described by the Zhang's formula [41]. The calibration of the spatial translation between the camera and projector was also critical for the accurate display of annotations.

The different display technology had a huge impact on the usability of the system. I had modified the Vuzix HMD for single eye use based on my preference, which was left eye dominance, when 70% of the population is right eye dominant, making the display unusable for the majority of other people.

A

APPENDIX

A.1 INERTIAL MEASUREMENT UNITS

Though work was done with the IMU to improve the tracking of SLAM algorithms, it was ultimately not successful and thus is not included in the main body of the thesis. The work done, as well as future considerations are detailed in this appendix.

Different approaches for pose estimation have different advantages and disadvantages as discussed in Section 2.5. Hybrid systems attempt to combine the beneficial characteristics of different approaches and compensate for the inherent weaknesses of each. We aimed to replicate the works of multiple authors [40, 11, 39] by incorporating accelerometers and gyroscopes.

An initial evaluation with a 9-Degree of Freedom (DOF) - Razor IMU was done. This IMU uses a ITG-3200 gyroscope[1], ADXL345 accelerometer[2], HMC5883L magnetometer[3]. All the sensors are mounted on a single board and the data is processed by an Atmel ATmega328[4]. The MPU-6050[5] with 6-DOF was also evaluated. All sensors and processors are included on a single chip and the information is output using the I^2C protocol. We used a FTDI VCP[6] chip to capture and process the information for PC. Because of the higher sample rate

1 http://invensense.com/mems/gyro/itg3200.html.
2 http://www.analog.com/en/mems-sensors/mems-inertial-sensors/adxl345/products/product.html
3 http://www.magneticsensors.com/three-axis-digital-compass.php
4 http://www.atmel.com/devices/atmega328.aspx
5 http://www.invensense.com/products/motion-tracking/6-axis/mpu-6050/
6 http://www.ftdichip.com/Products/ICs/FT200XD.html

and resolution we opted to use the MPU-6050 over the Razor. Below are the specifications for the MPU-6050:

- 1 KHz sample rate for gyroscope
- Programmable full scale accelerometer range of ± 2 g, ± 4 g, ± 8 g and ± 16 g
- Gyroscope sensitivity of up to 131 LSBs/dps
- Low power draw of 3.8 mA @ 3.4 V

The PTAM algorithm would occasionally lose tracking, especially with rapid movement. At the start of each new frame, the algorithm would estimate the transformation of the camera to begin a patch search for the previously detected features. When these estimates are wrong and the features lie outside of the search patch, the algorithm loses tracking.

To use the IMU data in pose tracking several steps were taken. First, the magnetometer along with pose captures form the camera calibration was used to calculate the rotation matrix between the camera and the axes of the IMU. Between frames, the rotational movement was calculated using the trapezoid rule to estimate the rotational of the camera. This rotation value was then used in the patch search algorithm for finding features.

The result of this was systematic errors in the value for rotation estimation which were unsuitable for use. This error could likely be due to many factors. The trapezoid estimation of the rotation was too coarse to be performed at 33 fps when the frames were captured. The work of integration of the gyroscope values should be delegated to the µcontroller to allow for much higher sample rates. The presence of noise in the magnetometer readings during calibration could have contributed to a bad calibration value. No filtering of the input

data was performed when the calibration took place. Due to time constraints, however, the reason for the failure was not fully investigated.

A.1.1 *IMU Calibrations*

We used static error analysis to determine the calibration values for the sensor. The sensor is held stationary and the sample rate is set to the highest value. The drift rate is then calculated in degrees per minute. The bias for the accelerometer and magnetometer is then calculated from averaging readings. The maximum and scale values for each of the sensors were also calculated due to variance in the manufacturing process. The magnetometer was also particularly sensitive to the presence of ferric objects in the vicinity and those objects should be removed.

The calibrations were specific to the devices we were using in this work and they are given here for record purposes and as a reminder that baseline values for all IMU devices need to be calculated for accurate readings.

Acceleration due to gravity Scale factors (min/max):

x −284/294
y −289/298
z −294/234

Magnetic field strength (min/max):[7]

x −150/294
y −202/206
z −254/200

Gyroscope drift:

7 This value will be dependent on the operating environment including the presence of heavy machinery etc. that may interfere with the magnetic field.

x 5.45
y 18.43
z 12.08

A.2 CAMERA CALIBRATION

A.2.1 *Kinect*

The Kinect camera gives the depth information as an 11 bit value. The raw sensor data can be converted to a depth range using Equation A.1. [26]. The information from the Kinect is acquired from the PrimeSense Drivers included with OpenNI [32, 27].

$$\mathrm{Range}(m) = \frac{1}{-0.0030711016 * raw\_depth + 3.3309495161}, \quad \text{(A.1)}$$

where raw_depth is the 11 bit value from the camera. This gives a reading in meters with an average of around 4% uncertainty and an optimal distance of about 2.5 m, where uncertainly drops to about 1% [5].

Using the calibration technique from Cámera Lúcida [30], the transformation for the projector's viewpoint compensation was calculated at:

$$\begin{pmatrix} -0.19 & -0.05 & 0.24 & -1.27754e^{-46} & 1.79851e^{-261} & -1.54897e^{-41} \end{pmatrix}$$

A.2.2 *Camera Calibration*

The RGB camera[8] used for the PTAM algorithm was calibrated using Zhang's pin-hole camera model [41]. This can be represented with k_1, k_2, k_3 as the radial distortion coefficients p_1, p_2 as tangential distortion coefficients.

$$\begin{pmatrix} 0.850797 & 1.13857 & 0.510602 & 0.488806 & 0.456279 \end{pmatrix}$$

8 https://www.microsoft.com/hardware/en-nz/p/lifecam-hd-3000/T3H-00014

A.3 CODE REPOSITORY

A repository of the code can be found at in the following address with revision number for commitments of interest for the project.

Svn://132.181.247.3/RARCollab

Commitment revisions of note are listed below with a description of the changes to the system.

- r19 Ability to annotate live video of PTAM from second window.
- r21 Ability to select keyframes and annotate from them. Ability to select keyframes and annotate from them
- r25 Keyframes drawn in color
- r26 Delaunay reconstruction for initial keyframes.
- r29 Delaunay reconstruction from multiple keyframes.
- r63 Networking code all working well PTAM split into two parts.
- r64 Laser Projector initial proposal.
- r76 Annotation using laser projector.
- r83 Servo Control of laser projector on a tripod.
- r94 Network code for laser projector.
- r101 Use of IMU with PTAM pose updating.
- r119 Added the ability for capture and calibration for the Collaboration tool.

A.4 HOUSING

The CAD drawings for the projector system are shown below. All parts are cut from 1mm sheet metal. The mounts allow Robotis Dynamixel DX-117 servo motors with HN01-N1 horns to be fixed to the housing.[9] The motors are controlled by a FT232R[10] FTDI USB to serial UART interface chip with the ability to set acceleration and movement profiles for the motors.

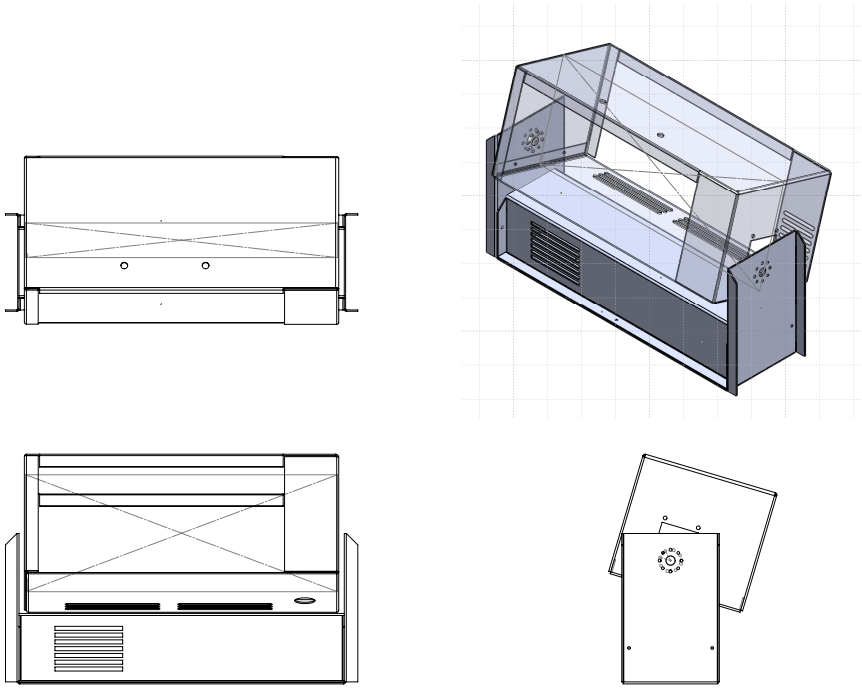


Figure 34: The final version of the assembled housing for the projector system. The projector cover is on top connected to the arm. The motor housing is between the two.

9 http://support.robotis.com/en/product/dynamixel/dx_series/dx-117.htm
10 www.ftdichip.com/Products/FT232R.htm

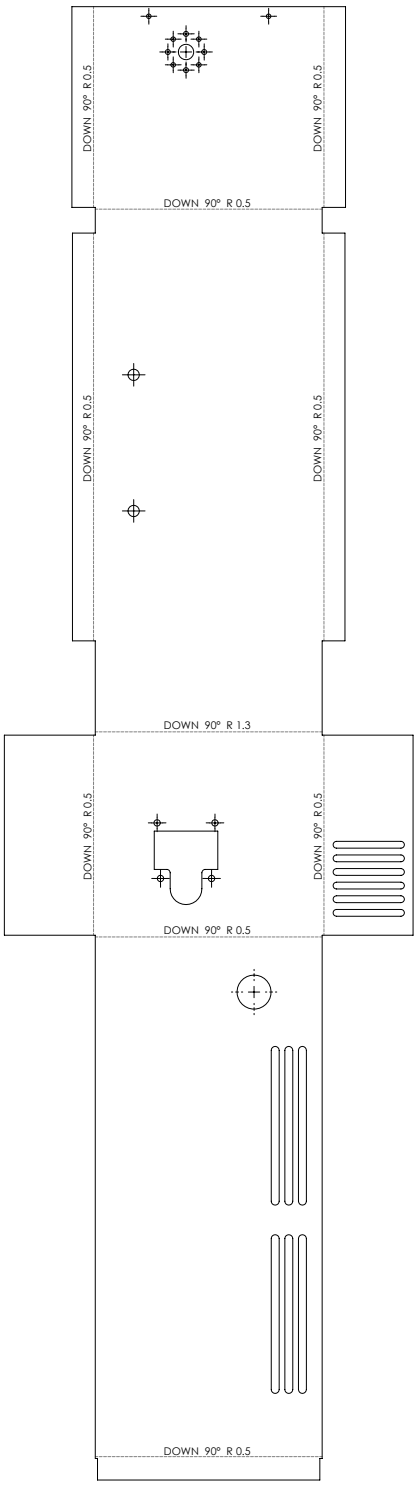


Figure 35: The Projector cover housing drawing and bend pattern. Scaled 1 : 5

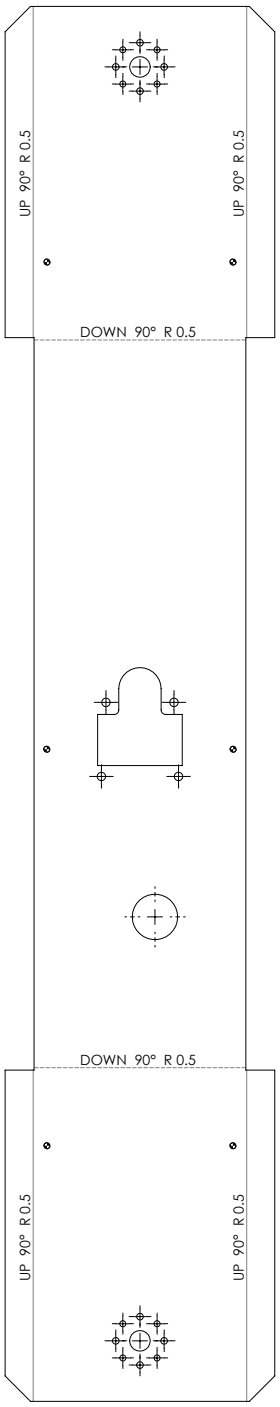


Figure 36: The arm drawing and bend pattern. Scaled 1 : 4

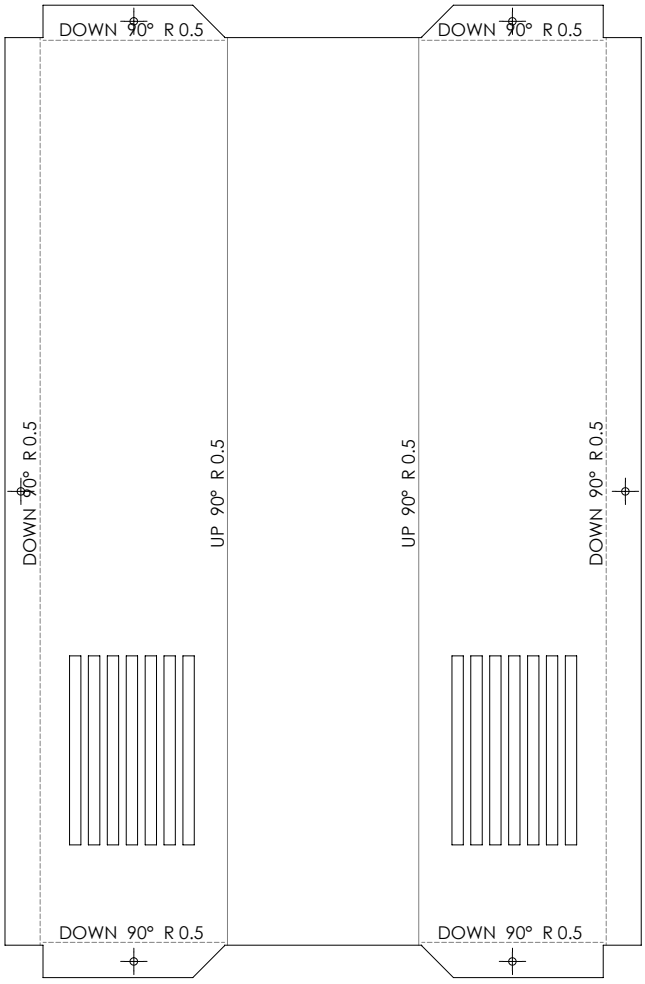


Figure 37: The motor housing drawing and bend pattern. Scaled 1 : 3

BIBLIOGRAPHY

[1] R. Grasse A. Duenser and M. Billinghurst. "A Survey of Evaluation Techniques Used in Augmented Reality Studies." In: *ACM SIGGRAPH Asia*. ACM, New York, USA, 2008.

[2] Mark Ashdown and Peter Robinson. "The Escritoire: Remote Collaboration in a Task Space." In: *Proceedings of the 2003 ACM SIGMM Workshop on Experiential Telepresence*. ETP '03. Berkeley, California: ACM, 2003, pp. 73–75. ISBN: 1-58113-775-3. DOI: 10.1145/982484.982498. URL: http://doi.acm.org/10.1145/982484.982498.

[3] Boost ASIO. *Boost ASIO Library*. May 2012. URL: http://www.boost.org/libs/asio/.

[4] M. Bauer, G. Kortuem, and Z. Segall. ""Where Are You Pointing At" A Study of Remote Collaboration in a Wearable Videoconference System." In: *Proceedings of the 3rd IEEE*. International Symposium on Wearable Computers. Oct. 1999, p. 155.

[5] Bas des Bouvrie. "Improving RGBD Indoor Mapping with IMU Data." MA thesis. Delft, the Netherlands: Delft University of Technology, 2011.

[6] CGAL, *Computational Geometry Algorithms Library*. URL: http://www.cgal.org.

[7] Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. "Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks." In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. CHI '03. New York, NY, USA: ACM, 2003, pp. 513–520. ISBN: 1-58113-630-7. DOI: 10.1145/642611.642701. URL: http://doi.acm.org/10.1145/642611.642701.

[8] Steffen Gauglitz et al. "In Touch with the Remote World: Remote Collaboration with Augmented Reality Drawings and Virtual Navigation." In: *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*. VRST '14. Edinburgh, Scotland: ACM, 2014, pp. 197–205. ISBN: 978-1-4503-3253-8. DOI: 10.1145/2671015.2671016. URL: http://doi.acm.org/10.1145/2671015.2671016.

[9] Pavel Gurevich et al. "TeleAdvisor: a versatile augmented reality tool for remote assistance." In: *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*. CHI '12. New York, NY, USA: ACM, 2012, pp. 619–622. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207763. URL: http://doi.acm.org/10.1145/2207676.2207763.

[10] K. Morinaga H. Kato M. Billinghurst and K. Tachibana. "The effect of spatial cues in augmented reality video conferencing." In: *9th International Conference on Human - Computer Interaction*. Lawrence Erlbaum Associates, NJ. New Orleans, pp. 478–481.

[11] Myung Hwangbo, Jun-Sik Kim, and T. Kanade. "Inertial-aided KLT feature tracking for a moving camera." In: *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. St. Louis, MO, USA: IEEE, Oct. 2009, pp. 1909–1916. ISBN: 978-1-4244-3803-7. DOI: 10.1109/IROS.2009.5354093. URL: http://dx.doi.org/10.1109/IROS.2009.5354093.

[12] Apple Inc. *PrimeSense 3D Sensor*. Mar. 2010. URL: http://www.primesense.com/files/FMF_2.PDF.

[13] H. Heibel P. Meier J . Platonov and B. Grollmann. "A Mobile Markerless AR System for Maintenance and Repair." In: *IEEE and ACM International Symposium on Mixed and Augmented Reality*. 5. Oct. 2006, pp. 105–108.

[14] Brett Jones et al. *Projection Mapping Central*. 2010. URL: http://projection-mapping.org/whatis/ (visited on 09/30/2012).

[15] O. Kahler et al. "Very High Frame Rate Volumetric Integration of Depth Images on Mobile Device." In: *IEEE Transactions on Vi-*

sualization and Computer Graphics (Proceedings International Symposium on Mixed and Augmented Reality 2015 22.11 (2015).

[16] K. Khoshelham. "Accuracy Analysis Of Kinect Depth Data." In: *ISPRS workshop laser scanning* 38 (Aug. 2011), pp. 5–7.

[17] Gab-Hoe Kim. "Vision-based simultaneous localization and mapping with two cameras." In: *Measurement* (2005), pp. 1671–1676. URL: http://ieeexplore.ieee.org.ez87.periodicos.capes.gov.br/stampPDF/getPDF.jsp?tp=&arnumber=1192152.

[18] Georg Klein and David Murray. "Parallel Tracking and Mapping for Small AR Workspaces." In: *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*. Nara, Japan, Nov. 2007.

[19] Takeshi Kurata et al. "Remote Collaboration using a Shoulder-Worn Active Camera/Laser." In: *Proceedings of the Eighth International Symposium on Wearable Computers*. ISWC '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 62–69. ISBN: 0-7695-2186-X. DOI: 10.1109/ISWC.2004.37. URL: http://dx.doi.org/10.1109/ISWC.2004.37.

[20] Stefan Leutenegger, Margarita Chli, and Roland Siegwart. "BRISK: Binary Robust Invariant Scalable Keypoints." In: *Proceedings of the IEEE International Conference on Computer Vision*. 2011.

[21] Mark Billinghurst and Andreas Duenser. *Project Title: Mobile AR for Hands-Free Remote Collaboration*. Tech. rep. The Hit Lab NZ, University of Canterbury, New Zealand, 2011.

[22] Ravi S. Menon et al. "Ocular Dominance in Human V1 Demonstrated by Functional Magnetic Resonance Imaging." In: *Journal of Neurophysiology* 77.5 (1997), pp. 2780–2787. ISSN: 0022-3077.

[23] Microsoft. *Kinect website*. Feb. 2012. URL: http://www.xbox.com:80/en.

[24] R.A. Newcombe, S. Lovegrove, and A.J. Davison. "DTAM: Dense tracking and mapping in real-time." In: *Proc. of the Intl. Conf. on Computer Vision (ICCV), Barcelona, Spain*. Vol. 1. 2011.

[25] Richard A. Newcombe et al. "KinectFusion: Real-time dense surface mapping and tracking." In: *ISMAR*. 2011, pp. 127–136.

[26] OpenKinect. *OpenKinect.org*. Feb. 2012. URL: http://openkinect.org.

[27] *OpenNI User Guide*. Last viewed 19-03-2012 11:32. OpenNI organization. Mar. 2012. URL: http://www.openni.org/documentation.

[28] Jiazhi Ou et al. "DOVE: drawing over video environment." In: *ACM Multimedia*. Ed. by Lawrence A. Rowe et al. ACM, 2003, pp. 100–101. ISBN: 1-58113-722-2. URL: http://dblp.uni-trier.de/db/conf/mm/mm2003.html#OuCFY03.

[29] Doug Palmer et al. "Annotating with light for remote guidance." In: *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces*. OZCHI '07. New York, NY, USA: ACM, 2007, pp. 103–110. ISBN: 978-1-59593-872-5. DOI: 10.1145/1324892.1324911. URL: http://doi.acm.org/10.1145/1324892.1324911.

[30] Christian Parsons. *Cámara Lúcida. R+D.* Last viewed 19-02-2012 12:44. Camara Lucida. 2012. URL: http://www.camara-lucida.com.ar/.

[31] Ronald Poelman et al. "As if Being There: Mediated Reality for Crime Scene Investigation." In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*. CSCW '12. New York, NY, USA: ACM, 2012, pp. 1267–1276. ISBN: 978-1-4503-1086-4. DOI: 10.1145/2145204.2145394. URL: http://doi.acm.org/10.1145/2145204.2145394.

[32] *Prime Sensor™ NITE 1.3 Algorithms notes*. Last viewed 19-02-2012 15:34. PrimeSense Inc. 2010. URL: http://www.primesense.com.

[33] Boost Serialization. *Boost Serialization Library*. May 2012. URL: http://www.boost.org/libs/serialization/.

[34] H. Stewénius, C. Engels, and D. Nistér. "Recent Developments on Direct Relative Orientation." In: *ISPRS Journal of Photogram-*

metry and Remote Sensing 60 (4 June 2006), pp. 284–294. URL: http://dx.doi.org/10.1016/j.isprsjprs.2006.03.005.

[35] *Structured Light vs Microsoft Kinect*. Revision 3.2. hackengineer. Mar. 2010. URL: http://www.hackengineer.com/structured-light-vs-microsoft-kinect/.

[36] Matthew Tait et al. "A Projected Augmented Reality System for Remote Collaboration." In: *ISMAR2013 - International Symposium on Mixed and Augmented Reality*. 2013.

[37] G. Tosolin et al. "From teaching machines to the exploitation of virtual and augmented reality: Behavior Analysis supporting manual workers in aerospace industry within the European project ManuVAR." In: *Proc. of the 3rd International Conference on Applied Human Factors and Ergonomics*. AHFE2010. Miami, Florida, USA: ACM, 2010, pp. 17–20. ISBN: 1-58113-775-3. DOI: 10.1145/982484.982498.

[38] Tony Tsai et al. "Method and Data Presenting Device For Assisting A Remote User To Provide Instructions." Patent PCT/EP2014/059558 (Europe). 2014.

[39] D.R. Wong, M.P. Hayes, and A. Bainbridge Smith. "IMU-aided SURF feature matching for relative pose estimation." In: *IVCNZ10*. 2010, pp. 1–6.

[40] Suya You and Ulrich Neumann. *Fusion of Vision and Gyro Tracking for Robust Augmented Reality Registration*. 2001.

[41] Zhengyou Zhang. "A Flexible New Technique for Camera Calibration." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.11 (Nov. 2000), pp. 1330–1334. ISSN: 0162-8828. DOI: 10.1109/34.888718. URL: http://dx.doi.org/10.1109/34.888718.

[42] Henrik Zimmer. "Voronoi and Delaunay techniques." In: *Proceedings of Lecture Notes, Computer Sciences* 8 (2005).

DECLARATION

I declare that this dissertation is my own work. It is being submitted for the degree of Master of Engineering at the University of Canterbury. It has not been submitted for any other degree or examination in any other University.

Christchurch, New Zealand, December 2015

Tony Tsai